ПРИКЛАДНАЯ СТАТИСТИКА

С.А. Айвазян
Б.М. Бухштабер
И.С. Енюков
Л.Д. Мешалкин

КЛАССИФИКАЦИЯ И СНИЖЕНИЕ РАЗМЕРНОСТИ

S. A. Aivazyan
V. M. Buchstaber
I. S. Yenyukov
L. D. Meshalkin

APPLIED STATISTICS

CLASSIFICATION AND REDUCTION OF DIMENSIONALITY

Reference edition

Edited by
prof. S. A. Aivazyan

Finansy i statistika
Moscow
1989

С. А. Айвазян
В. М. Бухштабер
И. С. Енюков
Л. Д. Мешалкин

ПРИКЛАДНАЯ СТАТИСТИКА

КЛАССИФИКАЦИЯ И СНИЖЕНИЕ РАЗМЕРНОСТИ

Справочное издание

Под редакцией
проф. С.А.Айвазяна

МОСКВА
"ФИНАНСЫ И СТАТИСТИКА"
1989

Рецензенты: *Б. Г. Миркин, Е. Г. Ясин*

П75 **Прикладная статистика: Классификация и снижение размерности:** Справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна.— М.: Финансы и статистика, 1989.— 607 с.: ил.
ISBN 5—279—00054—X.

Книга логически завершает справочные издания «Прикладная статистика: Основы моделирования и первичная обработка данных» (1983 г.) и «Прикладная статистика: Исследование зависимостей» (1985 г.). Рассматриваются задачи классификации объектов, снижения размерности. Большое внимание уделяется разведочному статистическому анализу.
Для специалистов, применяющих методы анализа данных

П $\frac{1702060000—036}{010(01)—89}$ 100—88 ББК 22.172

ISBN 5—279—00054—X

ПРЕДИСЛОВИЕ

Данная книга является третьей в трехтомном справочном издании, задуманном и реализуемом нашим авторским коллективом. В первом томе (Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983. — 472 с.) дается, в частности, определение *прикладной статистики* (см. с. 19) как самостоятельной научной дисциплины, разрабатывающей и систематизирующей понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов. Второй том (Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Исследование зависимостей. — М.: Финансы и статистика, 1985.— 488 с.) посвящен описанию методов анализа структуры, тесноты и конкретного вида статистических связей между исследуемыми признаками разной природы — количественными, ординальными, номинальными (категоризованными), а также обзору программного обеспечения этих методов. В числе описанных методов — корреляционный, регрессионный, дисперсионный, ковариационный анализ, элементы анализа временных рядов и систем одновременных эконометрических уравнений.

При минимальной вероятностно-статистической подготовке читателя, обеспечиваемой, например, обычным вероятностно-статистическим курсом экономического или технического вуза, *данный* (третий) *том пригоден для полностью автономного чтения* (т.е. его понимание не требует знания каких-либо специальных сведений, содержащихся в первых двух томах). Он посвящен актуальнейшим аспектам общей проблемы статистического анализа данных — *задачам классификации объектов, снижения размерности исследуемого признакового пространства и статистическим методам их*

решения. Лишь в последние два-три десятилетия, когда определенного уровня достигли вычислительная база исследований и теоретические разработки многомерного статистического анализа, главной проблемой теории и практики классификации и снижения размерности стало развитие достаточно изощренного и эффективного в приложениях математического аппарата. На этом пути уже имеются серьезные достижения, однако до сих пор в отечественной, да пожалуй, и в мировой специальной литературе не было издания, в котором эти достижения были бы достаточно полно просистематизированы, выстроены в общую методологическую схему, снабжены необходимыми практическими рекомендациями (включая вопросы преодоления вычислительных трудностей и использования подходящего типового программного обеспечения).

Авторы предлагаемой вниманию читателей книги ставили перед собой именно такую целевую установку. При этом изложение построено таким образом, что уже знакомство с «Введением» должно позволить читателю составить достаточно ясное представление о сущности, роли и назначении статистических методов классификации и снижения размерности, понять их разноаспектную типологизацию, узнать о содержании и логических связях всех частей книги (включая основные постановки задач и «адреса» их решений в книге). Следует отметить в общем замысле и в содержании книги один аспект, который выделяет ее среди другой литературы данного профиля. Речь идет о том специальном и неослабном внимании, которое уделяется в книге реализации важнейшего, узлового этапа всякого прикладного исследования, использующего математические методы и модели, — *этапа разведочного статистического анализа.* Как известно, назначение этого этапа — тщательный предварительный анализ, своеобразное «прощупывание» исходных статистических данных с целью выявления их вероятностной и геометрической природы, формирования и верификации тех или иных рабочих гипотез, касающихся этого аспекта проблемы. Принятые на этом этапе рабочие исходные допущения о математической модели реального механизма генерирования анализируемых данных являются определяющими в выборе необходимого математического инструментария, а значит, — и в успехе всего статистического исследования. Однако, к сожалению, в существующей практике прикладных статистических исследований этот важнейший этап чаще всего либо полностью игнорируется, либо реализуется весьма поверхностно. И одна из главных причин этого — почти полное отсутствие необходимой научно-методологи-

ческой литературы (изданный много лет назад перевод книги Дж. Тьюки «Разведочный анализ», в свое время весьма полезный, ныне приходится отнести к устаревшим источникам информации). В данной же книге эти вопросы занимают центральное место: так или иначе с ними связано большинство глав (кроме гл. 1—4), а *непосредственно* этой проблематике посвящен специальный раздел IV (гл. 18—21). Авторы старались сопровождать изложение этих важных вопросов подробным описанием существа, роли и научно-прикладного значения результатов, полученных отечественными специалистами (в сравнении с результатами зарубежных исследователей).

Книга состоит из 4 разделов и 21 главы.

Раздел I (гл. 1—4) посвящен задачам классификации в ситуации, когда исследователь обладает так называемыми обучающими выборками (т. е. «*классификации с учителем*»). Математический аппарат, используемый при решении подобных задач, объединяется в разделе многомерного статистического анализа, именуемого *дискриминантный анализ*.

Раздел II (гл. 5—12) посвящен задачам «*классификации без учителя*» (исследователь не располагает обучающими выборками). Математический аппарат решения таких задач включает в себя методы *кластер-анализа*, или *автоматической классификации* (в том числе иерархические процедуры классификации), а также статистические методы расщепления смесей вероятностных распределений.

Раздел III (гл. 13—17) содержит описание наиболее разработанных и эффективных методов снижения размерности исследуемого признакового пространства и отбора наиболее информативных показателей. Среди представленных здесь методов — главные компоненты, факторный анализ, метод экстремальной группировки параметров, многомерное шкалирование, экспертно-статистический метод построения интегрального (латентного) показателя, методы нелинейного отображения многомерных данных в пространства низкой размерности по различным критериям, анализ соответствий в случае неколичественных переменных.

Раздел IV (гл. 18—21) объединяет в себе описание методов так называемого разведочного статистического анализа и одновременно *вопросов вычислительной и программной реализации* представленных в книге методов, включая обзор по соответствующему программному обеспечению ЭВМ (в том числе *персональных* ЭВМ) и краткое освещение *проблем интеллектуализации статистического программного обеспечения*. Методы разведочного (предмодельного) статисти-

ческого анализа данных (и, в частности, методы целенаправленного проецирования многомерных наблюдений) направлены на «прощупывание» геометрической и вероятностной природы обрабатываемых данных с целью формирования адекватных реальности рабочих исходных допущений, на которых строится дальнейшее исследование. Эти методы как один из инструментов разведочного анализа являются естественным и необходимым дополнением к методам первичной статистической обработки, описанным в гл. 10, 11 первого тома данного издания. Сделанный в книге особый акцент на этих методах обусловлен тем обстоятельством, что в существовавшей до последнего времени практике статистических исследований этапу предмодельного анализа, методам выявления геометрической и вероятностной природы обрабатываемых данных, различным приемам тестирования гипотетических структур используемых моделей, как правило, не уделялось должного внимания.

В книгу включен ряд оригинальных результатов исследований авторов, а также результаты, ранее не публиковавшиеся в отечественной литературе: общая теория автоматической классификации (гл. 10), экспертно-статистический метод построения единого сводного показателя эффективности (гл. 15), некоторые приемы томографического анализа и целенаправленного проецирования многомерных данных (гл. 18—20), методы классификации при наличии элементов обучения (гл. 11), методы оцифровки неколичественных переменных (гл. 17).

Книга написана: С. А. Айвазяном — предисловие, введение, гл. 5 (без п. 5.4.7), 6, 13 (без § 13.6), 14, 15 и § 21.2; В. М. Бухштабером — гл. 7, 8, 10, 20, а также гл. 9, 19 (совместно с И. С. Енюковым); И. С. Енюковым — гл. 11, 12, 16, 17, 18, а также гл. 9, 19 (совместно с В. М. Бухштабером), § 13.6 и § 21.1; Л. Д. Мешалкиным — гл. 1, 2, 3, 4; п. 5.4.7 написан Б. Г. Миркиным.

Поскольку книга *завершает* труд коллектива авторов, посвященный кругу проблем, обозначенному как *прикладная статистика*, попробуем обсудить положение дел в этой области.

Подавляющее большинство исследователей и целых научных коллективов, работающих в области теории и практики статистического анализа данных, понимают и признают, что эффективность прикладной реализации математико-статистических методов, успешное развитие конкретных проблемно- и методо-ориентированных систем автоматизированной статистической обработки данных (представляю-

щих важную составную часть разнообразных автоматизированных систем поддержки принятия решений в различных отраслях человеческой деятельности) зависят не только от уровня *теоретических* вероятностно-статистических разработок (в этом плане отечественная школа традиционно относится к передовым), но и от степени продвинутости в разработке ряда смежных теоретических и прикладных проблем, остающихся, по существу, вне традиционных рамок *математической статистики*. И дело, разумеется, не в том, как именно назвать статистическую дисциплину, занимающуюся *комплексной разработкой всех необходимых инструментальных и методологических проблем*: в некоторых странах (во Франции, например) се чаще называют «а н а л и з о м д а н н ы х».

Мы считаем, что термин «прикладная статистика» вполне приемлем, тем более что он уже давно в обиходе в целом ряде стран (США, ФРГ и др.), в которых имеются специализации студентов, институты и журналы такого названия. Хотелось бы обратить внимание читателя на наиболее актуальные направления исследований этой научной дисциплины.

а) *Развитие методов анализа данных, не апеллирующих к их вероятностной природе, а также методов, нацеленных на выявление вероятностной и геометрической природы обрабатываемых данных в условиях отсутствия соответствующей априорной информации*. Именно таким методам уделено большое внимание в данной книге (кластер-анализ, многомерное шкалирование, томографические методы, целенаправленное проецирование многомерных данных и т. п., см. разделы II—IV книги) и именно они, как правило, оказываются вне поля зрения монографий и руководств по математической статистике.

б) *Формализация (математическая постановка) реальных задач статистического анализа данных в различных предметных областях (экономике, социологии, медицине и т. д.) и на базе этого опыта выработка типовых математических постановок задач*, выходящих за стеснительные рамки жестких канонических моделей. Этот самый важный и самый трудный этап математико-статистического исследования является и самым неблагодарным, поскольку de facto оказался как бы «незаконнорожденным дитем» теории и практики статистического анализа данных. Искусство реалистического моделирования формально не предусмотрено ни в одном из разделов инструментальной статистической науки, его развитие никак и ничем не стимулируется. Разрозненный положительный опыт такого рода, однако, при-

вел в последние полтора-два десятилетия к возникновению ряда интересных типовых постановок математических задач, связанных в основном с развитием подходов к получению *устойчивых* статистических выводов и к построению и обоснованию различных критериев качества метода, используемых в оптимизационных формулировках статистических задач (см. «Введение», а также § 2.4, 2.6, 3.1, 5.4, гл. 10, 15, 18 и др.).

в) *Вычислительные вопросы компьютерной реализации методов статистического анализа данных*. Это особенно актуально для сложных и подчас громоздких процедур многомерного статистического анализа. В книге этим вопросам посвящена (помимо отдельных пунктов) гл. 21.

г) *Теория и практика генерирования на ЭВМ данных заданной природы* и *развитие на этой основе методов статистического анализа малых выборок*. Этот подход предоставляет статистику эффективное (а иногда единственно возможное) средство исследования свойств обсуждаемых процедур многомерного статистического анализа, многие из которых не поддаются строгому аналитическому изучению.

д) *Развитие прикладного программного обеспечения по методам статистического анализа данных с акцентом на создание интеллектуализированных проблемно- и методо-ориентированных программных комплексов*, способных обеспечить исследователя развитой системой машинного ассистирования. В книге этим вопросам посвящена гл. 21.

Содержание и основные акценты теоретико-методологических и алгоритмических разработок прикладной статистики *гораздо динамичнее*, чем в традиционных математических дисциплинах, в том числе в *математической статистике*. Так, например, превалирующий удельный вес практической работы с существенно ограниченными выборками и возможности исследования свойств статистических процедур с помощью имитационного статистического моделирования на ЭВМ (см., например, г)) обусловливают исключение из категории актуальных (для *прикладной* статистики) значительной части асимптотической теории математической статистики. Равно как и теоретические разработки, основанные на понятии достаточности статистики или на принципе максимального правдоподобия, в той форме, как они формулируются сегодня: пока верили, что для распределений, близких к нормальному, целесообразно использовать рекомендации, гарантирующие оптимальность правил статистической обработки в рамках статистики нормального закона,

упомянутые понятия и подходы казались актуальными для приложений.

В свете сказанного нам представляется вполне оправданной и объективно назревшей необходимость специальных изданий по прикладной статистике.

Данное справочное издание адресовано как статистикам, экономистам, социологам, медикам и специалистам в других областях, использующим статистические методы классификации и снижения размерности в ходе решения задач, так и математикам, профессионалам-разработчикам описываемого математического аппарата (включая математиков-программистов). Специалист не математик может ограничиться «потребительским» стилем пользования данной книгой, при котором внимание сосредотачивается на постановках задач и рекомендациях по реализации предложенных решений (алгоритмах, описании диапазона их применимости, практических приемах анализа данных, программах), а усвоение обоснований этих рекомендаций и свойств используемых процедур не является необходимым.

В заключение одно важное, с нашей точки зрения, наблюдение. Все мы в настоящее время являемся свидетелями и в той или иной мере участниками набирающего все бóльшую силу глобального процесса информатизации общества. В проекции на проблематику данного издания это означает, в частности, что через сравнительно небольшое время персональный компьютер, а с ним и широкие возможности анализа данных станут неотъемлемой частью не только учрежденческого, но и домашнего уклада жизни. А следовательно, в повестке дня — бурная динамика роста спроса на методы и программы прикладной статистики.

Научный и научно-методический багаж, послуживший основой для написания данного издания, наработан авторами в основном в рамках их деятельности в Московском государственном университете им. М. В. Ломоносова, в Центральном экономико-математическом институте АН СССР, в Главном научно-исследовательском вычислительном центре 4-го Главного управления при Министерстве здравоохранения СССР и во Всесоюзном научно-исследовательском институте физико-технических и радиотехнических измерений Госстандарта СССР.

Бесспорное влияние на замысел и содержание книги оказали постоянные контакты авторов со своими коллегами по научному семинару «Многомерный статистический анализ и вероятностное моделирование реальных процессов», дейст-

вующему с 1969 г. в рамках Научного совета АН СССР по комплексной проблеме «Оптимальное планирование и управление народным хозяйством» и Совета по автоматизации научных исследований при Президиуме АН СССР, а также по Всесоюзному научно-методическому семинару «Вычислительные вопросы математической статистики», функционирующему в Московском государственном университете им. М. В. Ломоносова под руководством акад. Ю. В. Прохорова.

Авторы признательны Е. Г. Ясину и Б. Г. Миркину, взявшим на себя труд прочесть рукопись настоящего издания и сделавшим ряд ценных замечаний.

С. А. Айвазян

ВВЕДЕНИЕ. КЛАССИФИКАЦИЯ И СНИЖЕНИЕ РАЗМЕРНОСТИ. СУЩНОСТЬ И ТИПОЛОГИЗАЦИЯ ЗАДАЧ, ОБЛАСТИ ПРИМЕНЕНИЯ

В.1. Сущность задач классификации и снижения размерности и некоторые базовые идеи аппарата многомерного статистического анализа

Необходимость анализа и формализации задач, связанных со сравнением и классификацией объектов, сознавали ученые далекого прошлого. «Его (Аристотеля) величайшим и в то же время чреватым наиболее опасными последствиями вкладом в науку была идея классификации, которая проходит через все его работы ... Аристотель ввел или, по крайней мере, кодифицировал способ классификации предметов, основанный на сходстве и различии ...», — писал Дж. Бернал в «Науке истории общества» (М.: Изд-во иностр. лит., 1956.— С. 117). После Аристотеля с его «деревом вещей жизни» имеется еще в докомпьютерной эре ряд интереснейших примеров прекрасно построенных классификаций как в естественных, так и в общественных науках. Упомянем здесь (в хронологическом порядке) три из них: а) иерархическая классификация (основанная на понятии сходства) растений и видов М. Адансона (1757 г., [170]); б) знаменитая периодическая система элементов Д. И. Менделеева (1869 г.), представляющая собой, по существу, классификацию многомерных наблюдений (каждый химический элемент может быть представлен в виде вектора характеризующих его разнотипных признаков, включая характеристики конфигурации внешних электронных оболочек атомов) с выявленным единым классифицирующим фактором (зарядом атомного ядра) и с упорядочением элементов внутри каждого класса; в) классификация крестьянских хозяйств, уездов и губерний России по характеру и уровню развития капитализма, полученная В. И. Лениным на основе анализа земско-статистических подворных переписей (1899 г., [1]).

Надо сказать, что, хотя авторы упомянутых выдающихся классификаций и не располагали современным математическим аппаратом многомерного статистического анализа, основные *идеи* и *методологические принципы* этого аппарата явно или неявно пронизывают логику их конструкций, а

подчас (в частности, в работе В.И. Ленина «Развитие капитализма в России») и прямо формулируются.

Остановимся на четырех генеральных идеях и методологических принципах многомерного статистического анализа, на которых базируются, по существу, все основные разделы и подходы математического аппарата классификации и снижения размерности.

1. **Эффект существенной многомерности.** Сущность этого принципа в том, что выводы, получаемые в результате анализа и классификации множества статистически обследованных (по ряду свойств) объектов, должны опираться *одновременно на совокупность этих взаимосвязанных свойств с обязательным учетом структуры и характера их связей*. В [5] природа эффекта существенной многомерности поясняется на таком примере: попытка различить два типа потребительского поведения семей, основанная на последовательном применении критерия однородности Стьюдента [12, п. 11.2.8] сначала по одному признаку (удельные расходы на питание), потом по другому (удельные расходы на промышленные товары и услуги) не дала результата, в то время как *многомерный* аналог этого критерия [12, п. 11.2.9], основанный на так называемом расстоянии Махаланобиса и учитывающий *одновременно значения обоих упомянутых признаков и характер статистической связи между ними*, дает правильный результат (т. е. обнаруживает статистически значимое различие между двумя анализируемыми совокупностями семей). Формулировку существа этого принципа мы находим уже в упомянутой работе В. И. Ленина [1]. Возражая против классификации крестьянских хозяйств изолированно по каждому из анализируемых признаков с ориентацией на их средние значения, он пишет [1, с. 96]: «Признаки для различения этих типов должны быть взяты сообразно с местными условиями и формами земледелия; если при экстенсивном зерновом хозяйстве можно ограничиться группировкой по посеву (или по рабочему скоту), то при других условиях необходимо принять в расчет посев промышленных растений, техническую обработку сельскохозяйственных продуктов, посев корнеплодов или кормовых трав, молочное хозяйство, огородничество и т. д. *Когда крестьянство соединяет в широких размерах и земледельческие и промысловые занятия, — необходима комбинация двух указанных систем группировки* (курсив наш. — *С. А.*), т. е. группировки по размерам и типам земледелия и группировки по размерам и типам «промыслов». Вопрос о приемах сводки подворных записей о крестьянском хозяйстве вовсе не такой узко специальный и второстепенный

вопрос ... Вследствие неудовлетворительной сводки масса драгоценнейших сведений прямо-таки теряется, и исследователь получает в свое распоряжение только «средние» цифры (по общинам, волостям, разрядам крестьян, по величине надела и т. д.). А эти «средние» ... зачастую совершенно фиктивны». Итак, статистический анализ множества объектов, даже если по каждому из них зарегистрированы значения набора признаков, будет неполным, ущербным, если ограничиваться при этом только средними значениями признаков и не использовать разнообразные характеристики тесноты и структуры связей между ними.

2. Возможность лаконичного объяснения природы анализируемых многомерных структур. Определим вначале, что понимается (здесь и в дальнейшем изложении) под *многомерной структурой*. Речь идет о множестве статистически обследованных объектов $\{O_1, O_2, \ldots, O_n\}$. Результаты статистического обследования представляются, как правило (но не всегда), в одной из двух форм:

таблицы (матрицы) «объект — свойство» вида

$$\mathbf{X} = (X_1, X_2, \ldots, X_n), \tag{B.1}$$

в которой $X_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)})'$ — вектор значений[1] анализируемых признаков (свойств) $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, зарегистрированных на i-м обследованном объекте;

или матрицы (таблицы) попарных сравнений вида

$$\mathbf{A} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ a_{21} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}, \tag{B.1'}$$

где элемент a_{ij} определяет результат сопоставления объектов O_i и O_j в смысле некоторого заданного отношения: a_{ij} может выражать меру сходства или различия объектов O_i и O_j, меру их связи или взаимодействия в каком-либо процессе (например, поток продукции отрасли i в отрасль j), геометрическое расстояние между объектами, отношение предпочтения ($a_{ij} = 1$, если объект O_i не хуже объекта O_j, и $a_{ij} = 0$ в противном случае) и т. д. Под возможностью лаконичного объяснения природы анализируемой многомерной струк-

---

[1] Штрих как верхний индекс матрицы или вектора означает операцию их транспонирования.

туры подразумевается априорное допущение, в соответствии с которым существует небольшое (в сравнении с p) число определяющих (типообразующих) факторов, с помощью которых могут быть достаточно точно описаны как наблюдаемые характеристики анализируемых объектов (т. е. все элементы $x_i^{(k)}$ и a_{ij} соответственно матриц **X** и **A**) и характер связей между ними, так и искомая классификация самих объектов. При этом упомянутые определяющие факторы могут находиться среди статистически обследованных характеристик, а могут быть *латентными*, т.е. непосредственно статистически не наблюдаемыми, но восстанавливаемыми по исходным данным вида (В.1) или (В.1′). Гениальный пример практической реализации этого принципа дает нам периодическая система элементов Менделеева: в этом случае роль идеально информативного единственного определяющего фактора играет, как известно, заряд атомного ядра элемента.

Отметим, что на данном принципе многомерного статистического анализа построены такие важнейшие разделы математического аппарата классификации и снижения размерности, как метод главных компонент и факторный анализ (см. гл. 13, 14), многомерное шкалирование (см. гл. 16), целенаправленное проецирование в разведочном статистическом анализе данных (см. гл. 18—21) и др.

3. Максимальное использование «обучения» в настройке математических моделей классификации и снижения размерности. Для пояснения этого принципа представим задачи классификации и снижения размерности по схеме «на входе задачи — на выходе задачи» (табл. В.1).

Таблица В.1

Объект	На «входе» задачи классификации и снижения размерности	На «выходе» задачи классификации	На «выходе» задачи снижения размерности
O_i $(i=1,2, \ldots, n)$	$\left(x_i^{(1)}, \ldots, x_i^{(p)}\right)$ или $(a_{i1}, \ldots, a_{in})$ (см. ф-лы (В 1) и (В 1′))	Номер класса, к которому отнесен O_i, или перечень статистически представленных в форме (В 1) или (В 1′) объектов, входящих в тот же класс, что и объект O_i	Значения искомых определяющих (типообразующих) факторов, характеризующих объект O_i

Если исследователь располагает и «входами» и «выходами» задачи, то исходную информацию называют *обучающей* и целью исследования является описание процедур, с помощью которых при поступлении только входных данных нового объекта его можно было бы с наибольшей (в определенном смысле) точностью отнести к одному из классов (в задаче классификации) или снабдить значениями определяющих факторов (в задаче снижения размерности). Именно к таким ситуациям относятся типичные задачи медицинской диагностики, когда в клинических условиях в качестве исходных данных исследователь располагает как «входами»— результатами инструментальных обследований пациентов, так и «выходами» — уже установленным диагнозом («болен»— «здоров») по каждому из них. Цель исследований такого типа — использование имеющегося «обучения» для отбора из множества результатов обследований небольшого числа наиболее информативных (с точки зрения диагностической силы) показателей и для построения на их основе формального диагностирующего правила (см., например, [115]).

Однако в задачах социально-экономического профиля исследователь, как правило, располагает в качестве исходных данных лишь «входной» информацией (второй столбец табл. В.1) и в лучшем случае отдельными элементами «обучения»: например, известно, что определенная группа (из числа статистически обследованных) объектов относится к одному и тому же классу, но какие есть другие классы и как между ними распределены остальные статистически обследованные объекты, неизвестно. Сущность обсуждаемого принципа как раз и состоит в том, что даже такая урезанная и обедненная обучающая информация оказывается весьма полезной в решении узловых задач «настройки» используемых математических моделей, как, например, выбор метрики в исследуемом признаковом пространстве, оценка общего числа классов, выбор критерия качества классификации и т. д.

4. Оптимизационная формулировка задач классификации и снижения размерности. Среди множества возможных методов, реализующих поставленную цель статистической обработки данных (разбиение совокупности статистически обследованных объектов на однородные классы, переход от заданного широкого набора признаков $x^{(1)},\ldots, x^{(p)}$ к небольшому числу определяющих факторов), нужно уметь найти *наилучший* метод с помощью оптимизации некоторого экзогенно заданного критерия (функционала) качества метода. Выбор конкретного вида этого критерия основан либо на априорном знании вероятностной и геометрической природы

обрабатываемых данных, либо на соображениях содержательного (экономического, медицинского, технического и т. п.) плана. В сочетании с некоторыми другими (более специфицированными) базовыми идеями [1] этот подход дает возможность построить достаточно общую математическую конструкцию, в рамках которой удается «навести порядок» в огромном множестве существующих алгоритмов классификации и снижения размерности, подчас стихийно (и эвристически) возникающих из нужд разнообразных приложений.

В.2. Типовые задачи практики и конечные прикладные цели исследований, использующих методы классификации и снижения размерности

До разработки аппарата многомерного статистического анализа и, главное, до появления и развития достаточно мощной электронно-вычислительной базы главные проблемы теории и практики классификации и снижения размерности относились не к разработке методов и алгоритмов, а к полноте и тщательности отбора и теоретического анализа изучаемых объектов, характеризующих их признаков, смысла и числа градаций по каждому из этих признаков.

Все методы классификации сводились, по существу, к методу так называемой *комбинационной группировки*, когда все характеризующие объект признаки носят дискретный характер или сводятся к таковым (пол или мотив миграции индивидуума, уровень жилищных условий или число детей в семье и т.п.), а два объекта относятся к одной группе только при точном совпадении зарегистрированных на них градаций одновременно по всем характеризующим их признакам (одинаковый пол, мотив миграции и т. д.). Методы снижения размерности ограничивались простым агрегированием однотипных признаков (например, переход от фиксации семейных расходов отдельно на молоко, сыр, сметану и т.п. к общим семейным расходам на молочные продукты) и отбором (на уровне содержательного анализа) некоторой на-

[1] Например, идея расширительного толкования понятия ядра класса ядром класса может быть точка, группа точек, ось, поверхность, случайная переменная и т д На этом, в частности, построен весьма общий подход к решению задач анализа данных, названный авторами «методом динамических сгущений» [106] Эта же идея использована нами и при построении общей теории автоматической классификации (гл 10).

иболее информативной части из исходного набора признаков.

Однако по мере роста объемов перерабатываемой информации и, в частности, числа классифицируемых объектов и характеризующих их признаков возможность эффективной реализации подобной логики исследования становилась все менее реальной (так, например, число k групп или классов, подсчитываемое при комбинационной группировке по формуле $k = m_1 \cdot m_2 \ldots m_p$, где m_j — число градаций по признаку $x^{(j)}$, а p — общее число анализируемых признаков, уже при $m_j = 3$ и $p = 5$ оказывается равным 243). Именно электронно-вычислительная техника стала тем главным инструментом, который позволил по-новому подойти к решению этой важной проблемы и, в частности, конструктивно воспользоваться разработанным к этому времени мощным аппаратом многомерного статистического анализа: методами распознавания образов «с учителем» (дискриминантный анализ) и «без учителя» (автоматическая классификация, или кластер-анализ), методами и моделями факторного анализа, многомерного шкалирования и т. д

Развитие электронно-вычислительной техники как средства обработки больших массивов данных стимулировало проведение в последние годы широких комплексных исследований сложных социально-экономических, технических, медицинских и других процессов и систем, таких, как образ и уровень жизни населения, совершенствование организационных систем, региональная дифференциация социально-экономического развития, планирование и прогнозирование отраслевых систем, закономерности возникновения сбоев (в технике) или заболеваний (в медицине) и т п. В связи с многоплановостью и сложностью этих объектов и процессов данные о них по необходимости носят *многомерный* и *разнотипный* характер, так как до их анализа обычно бывает неясно, насколько существенно то или иное свойство для конкретной цели. В этих условиях выходят на первый план проблемы построения группировок и классификации по многомерным данным (т е. проблемы *классификации многомерных наблюдений*), причем появляется возможность оптимизации этого построения с точки зрения наибольшего соответствия получаемого результата поставленной конечной цели классификации.

Цели классификации существенно расширяются, и одновременно содержание самого процесса классификации становится неизмеримо богаче и сложнее. Оно, в частности, дополняется *проблемой построения самой процедуры классификации*, ранее носившей чисто технический характер.

Для пояснения сущности основных типов задач классификации и конечных прикладных целей, которые ставит при этом перед собой исследователь, рассмотрим примеры.

Пример В.1. Выявление типологии потребительского поведения населения, анализ сущности дифференциации этого поведения, прогноз структуры потребления [154].

В качестве исходной информационной базы используются данные бюджетных обследований семей [105]. Поясним логическую схему исследования. Многомерная статистика рассматривает совокупность изучаемых многомерных объектов (В.1) как совокупность точек или векторов в пространстве описывающих их признаков. Применительно к схеме потребления совокупностью объектов, подлежащих изучению, является множество элементарных потребительских ячеек — семей. Каждая семья характеризуется, с одной стороны, некоторым набором X *факторов-детерминантов* (социально-демографические и другие признаки, описывающие условия жизнедеятельности семьи), а с другой — набором Y параметров поведения («переменных поведения»), в которых отражаются ее фактические потребности.

В качестве социально-демографических факторов, имеющих существенное значение для изучения потребительских аспектов социальной жизни, целесообразно использовать, например, общественную и национальную принадлежность, уровень образования и квалификацию, характер труда, демографический тип и возраст семьи, тип населенного пункта и характер жилища, размер и структуру имущества, уровень доходов.

Имеется некоторое сомнение относительно включения последнего показателя (уровень доходов), так как принципиально он может быть выражен через другие социально-демографические характеристики. Величина доходов является производной от уровня образования, квалификации, характера трудовой деятельности (через заработки работников семьи), половозрастного и численного состава семьи [1]. Поэтому доход остается в нашей конструкции как один из вспомогательных компонентов, в концентрированном виде выражающий разницу в основных факторах-детерминантах.

Различия в потребностях, складывающиеся под влиянием социально-демографических и природно-климатических

[1] Рассматривая в дифференцированном балансе доходов и потребления населения структуру потребления семей только по признаку различий в доходе, мы фактически абстрагируемся от всех других социально-демографических факторов.

условий, являются объективно существующими; они формируют весь строй поведения потребителя в конкретно-исторических условиях, а в конечном счете порождают своеобразные типы потребителей, ориентированные на существенно разное потребление.

Весь комплекс социально-демографических и других факторов, существенно воздействующих на структуру потребления, будем называть *типообразующим*. Они имеют определяющее значение, в то время как все другие дают лишь случайную вариацию в пределах одной группы (типа) потребительского поведения.

В качестве признаков поведения Y можно рассматривать три группы параметров: а) уровень и структуру потребления; б) характер (объем и содержание) использования свободного времени; в) интенсивность изменения социального, трудового, демографического статуса (в [154] рассмотрена только первая группа признаков).

Итак, в задаче даны числовые характеристики и градации типообразующих и одновременно поведенческих признаков каждой семьи из анализируемой совокупности.

Решение общей проблемы, связанной с выявлением и прогнозом структуры и дифференциации потребностей населения, распадается в соответствии с принятой в [154] логической схемой исследования на следующие этапы.

1. *Сбор и первичная статистическая обработка исходных данных*. Исследуемые объекты (семьи) выступают в качестве многомерных наблюдений или точек в двух многомерных пространствах признаков. Фиксируя в качестве координат этих точек значения (или градации) типообразующих переменных X (т. е факторов-детерминантов), рассматриваем их в «пространстве состояния» $\Pi(X)$, т. е. в пространстве, координатами которого служат основные показатели жизнедеятельности семей. Фиксируя же в качестве координат тех же самых объектов значения показателей Y их потребительского поведения, рассматриваем их в «пространстве поведения» $\Pi(Y)$. Очевидно, при надлежащем выборе метрики в пространствах $\Pi(X)$ и $\Pi(Y)$ геометрическая близость двух точек в $\Pi(X)$ будет означать сходство условий жизнедеятельности соответствующих двух семей, так же как и геометрическая близость точек в $\Pi(Y)$ будет означать сходство их потребительского поведения. Среди методов первичной статистической обработки анализируемых данных, обычно используемых на этой стадии исследования (см., например, [12, гл. 10—11]), широко распространенными и весьма полезными являются методы изучения различных одно-, двух- и трехмерных эмпирических распределе-

ний, которые сводятся к построению и различным представлениям (графическим, табличным) упомянутых выше комбинационных группировок. Пример табличного представления одной из таких двумерных комбинационных группировок приведен в табл. В.2.

Таблица В.2

Градации признака $x^{(1)}$ (доход)	Градации признака $x^{(2)}$ (жилищные условия)				Сумма
	«низкое»	«удовлетворительное»	«хорошее»	«очень хорошее»	
«Низкий»	24	12	4	0	40
«Средний»	20	100	140	20	280
«Высокий»	4	8	28	40	80
Сумма	48	120	172	60	400

Эта комбинационная группировка построена на основе статистического обследования 400 семей по двум признакам из пространства $\Pi(X)$: по $x^{(1)}$ (руб.) — величине среднедушевого семейного дохода (с тремя градациями: «низкий», «средний» и «высокий»), и по $x^{(2)}$ — качеству жилищных условий (с четырьмя градациями: «низкое», «удовлетворительное», «хорошее» и «очень хорошее»). Каждая клетка таблицы соответствует классу, полученному в результате проведенной комбинационной группировки; внутри клетки обозначено число семей, имеющих данное сочетание градаций анализируемых признаков (подобные таблицы называют также «таблицами сопряженности», см., например, [12, п. 10.3.5], а также [11, 3.1]).

Для более полного представления результатов подобной классификации можно было бы ввести в программу компьютера требование выпечатывать номера семей, попавших в каждую из двадцати клеточек таблицы.

Заметим, что *непрерывным аналогом комбинационной группировки* является обычный переход от исходных наблюдений непрерывной случайной величины к «группированным» выборочным данным [12, п. 5.4.2]. Результат такого перехода представляется либо в виде таблицы, подобной табл. В.2, либо в виде графика (*гистограммы*).

2. *Выявление основных типов потребления с помощью разбиения исследуемого множества точек-семей на классы в «пространстве поведения»* $\Pi(Y)$. Гипотеза существова-

ния «естественных», объективно обусловленных типов поведения, т. е. какого-то небольшого количества классов семей, таких, что семьи одного класса характеризуются сравнительно сходным, однотипным потребительским поведением, геометрически означает распадение исследуемой в «пространстве поведения» совокупности точек-семей на соответствующее число «сгустков» или «скоплений» точек. Выявив с помощью подходящих методов многомерного статистического анализа (кластер-анализа, таксономии) эти классы-сгустки, тем самым определим основные типы потребительского поведения. Попутно в качестве «побочного результата» решения главной задачи этого этапа конструктивно реализуется метод построения целевых функций предпочтения, являющийся развитием и некоторой модификацией метода, предложенного в [47]. По существу, при этом решается одна и та же задача *регрессионного анализа* [11], но функция регрессии строится *отдельно только по однородным данным, попавшим в один какой-то класс.*

3. *Отбор наиболее информативных типообразующих признаков (факторов-детерминантов) и выбор метрики в пространстве типообразующих признаков.* Очевидно, неправомерно рассчитывать на то, что диапазоны возможных значений каждого из кандидатов в типообразующие признаки окажутся непересекающимися для семей с разным типом потребительского поведения. Другими словами, значения каждого из признаков $x^{(j)}$ в отдельности и их набора в совокупности подвержены некоторому неконтролируемому разбросу при анализе семей внутри каждого из типов потребления. Естественно считать наиболее информативными те факторы-детерминанты или те их наборы, разница в законах распределения которых оказывается наибольшей при переходе от одного класса потребительского поведения к другому. Эта идея и положена в основу метода отбора наиболее информативных (типообразующих) признаков-детерминантов. Наконец, отобрав небольшое число наиболее информативных признаков-детерминантов, мы можем попытаться снова разбить исследуемую совокупность семей на классы-сгустки, но уже в пространстве выявленных типообразующих признаков. При этом результат разбиения будет существенно зависеть не только от состава группы наиболее информативных типообразующих признаков, но и от способа вычисления расстояния между двумя точками-семьями в этом пространстве и, в частности, от того, с какими весами участвуют в этом расстоянии отобранные типообразующие признаки. Поэтому веса подбираются таким образом, чтобы результат разбиения семей на классы в про-

странстве наиболее информативных факторов-детерминантов в некотором смысле наименее отличался бы от разбиения тех же точек-семей, которое было получено в «пространстве поведения». Таким образом, добиваемся наибольшего совпадения, *наибольшей связности* в результатах классификации одного и того же множества семей в двух разных признаковых пространствах — «пространстве поведения» $\Pi(Y)$ и пространстве типообразующих признаков $\Pi(X)$.

4. *Анализ динамики структуры исследуемой совокупности семей в пространстве наиболее информативных типообразующих признаков.* Конечной целью этого этапа является прогноз тех постепенных преобразований классификационной структуры совокупности потребителей (семей, рассматриваемых в пространстве типообразующих признаков), которые должны произойти с течением времени. Реализация этапа может быть осуществлена с использованием результатов и подходов, описанных в [50], а также с помощью привлечения математического аппарата марковских цепей (аналогично тому, как используется этот аппарат при анализе динамики структуры трудовых ресурсов; см., например, [17]) и многомерных временных рядов [146]. При этом, конечно, должны быть учтены существующие методы прогноза социально-демографической структуры населения [31], [145].

5. *Прогноз структуры потребления.* На этом этапе исследования опираемся на результаты, полученные в итоге проведения предыдущего этапа, т. е. исходим из заданной классификационной структуры потребителей в интересующий нас период времени в будущем. Восстанавливая классификационную структуру потребления (классификационную структуру совокупности семей в пространстве признаков $\Pi(Y)$, характеризующих потребительское поведение семьи) по классификационной структуре потребителей (по классификационной структуре той же совокупности, но в пространстве типообразующих признаков), будем относить каждую конкретную семью к тому типу потребления, для которого значения характеризующих ее типообразующих признаков являются, грубо говоря, наиболее типичными.

Пример В. 2. Классификация как необходимый предварительный этап статистической обработки многомерных данных [9]. Пусть исследуется зависимость интенсивности миграции населения $x^{(p)}$ (профессиональной или территориальной) от ряда социально-экономических и географических факторов $x^{(1)}, x^{(2)}, \ldots, x^{(p-1)}$, таких, как средний заработок, обеспеченность жилой площадью, детскими учреждениями, уровень образования, возможности профессиональ-

ного роста, климатические условия и т. п. Естественно предположить (и результаты исследования это подтверждают), что для различных однородных групп индивидуумов одни и те же факторы влияют на $x^{(p)}$ в разной степени, а иногда и в противоположных направлениях. Поэтому до применения аппарата регрессионно-корреляционного анализа следует разбить все имеющиеся в нашем распоряжении данные $X_i' = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)})$ $(i = 1, 2, \ldots, n)$ на однородные классы и *решать далее поставленную задачу отдельно для каждого такого класса.* Только в этом случае можно ожидать, что полученные коэффициенты регрессии $x^{(p)}$ по $x^{(1)}$, $x^{(2)}$, ..., ..., $x^{(p-1)}$ будут допускать содержательную интерпретацию, а мера тесноты связи между $x^{(p)}$ и $(x^{(1)}, x^{(2)}, \ldots, x^{(p-1)})$ окажется достаточно высокой. Подобные задачи можно найти в [130, с. 77].

Другой вариант такого рода примера получим, если в качестве объектов исследования рассмотрим предприятия определенной отрасли, а в качестве вектора наблюдений X_i — совокупность объективных (нерегулируемых) условий работы i-го обследованного предприятия (сырье, энергия, оснащенность техникой и рабочей силой и т. п.). Классификация предприятий по X производится как необходимый предварительный этап для возможности последующей объективной оценки работы коллективов и разработки обоснованных дифференцированных нормативов: очевидно, лишь к предприятиям, попавшим в один класс по X, может быть применена одинаковая система нормативов и стимулирующих показателей (см. описание подобной задачи в [53]). Далее можно рассматривать задачу, аналогичную сформулированной выше, а именно: если $Y = (y^{(1)}, \ldots, y^{(q)})'$ — вектор показателей качества работы предприятия (объем и качество выпускаемой продукции, ее себестоимость, рентабельность и т.п.), а $U = (u^{(1)}, \ldots, u^{(m)})'$ — вектор регулируемых факторов, от которых зависят условия производства (число основных подразделений, уровень автоматизации и т. д.), то задачу описания интересующей нас зависимости вида $Y = f(U)$ естественно решать отдельно для каждого класса по X.

Пример В.3. Классификация в задачах планирования выборочных обследований [9]. Здесь речь пойдет о планировании выборочных экономико-социологических обследований городов. Предположим, что необходимо достаточно детально проанализировать подробные статистические данные о городах с целью выявления наиболее характерных черт в экономико-социологическом облике типичного среднерусского города. Производить подробный, кропотливый

анализ по каждому из городов РСФСР, очевидно, слишком трудоемко, да и нецелесообразно. По-видимому, разумнее попытаться предварительно выявить число и состав различных типов в совокупности обследованных городов по набору достаточно агрегированных признаков ($x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$), характеризующих каждый город (например, понимать под $x^{(j)}$ число жителей города, приходящееся на каждую тысячу жителей, обладающих заданным j-м признаком, скажем, высшим образованием, специальностью металлурга и т.п.). А затем, отметив наиболее типичные города в каждом классе (наблюдения-точки X_i, наиболее близко располагающиеся к «центрам тяжести» своих классов), отобрать их для дальнейшего (более детализованного) социально-экономического анализа. При этом, очевидно, мера представительности отобранных «типичных городов» определится удельным весом количественного состава точек данного класса среди всех рассматриваемых точек (городов). Подробнее об этой задаче см. в [9, гл. 5]. Похожие задачи планирования выборочных обследований с использованием методов классификации рассмотрены в [130, с. 34].

Анализ рассмотренных примеров с учетом, конечно, и другого накопившегося к настоящему времени опыта решения практических задач классификации в экономике, социологии, психологии, технике, медицине, геологии, археологии и других сферах практической и научной деятельности человека позволяет произвести определенную систематизацию этих задач в соответствии с их основными типами и конечными прикладными целями исследования (табл. В.3).

В качестве комментария к табл. В.3 поясним методологическую общность задач 3.1—3.3: прогноза экономико-социологических ситуаций, диагностики и автоматического распознавания зрительных и слуховых образов. Для этого лежащую в основе их решения методологическую схему связной неупорядоченной типологизации представим следующим образом. Пусть в качестве исходных данных об объекте O_i ($i = 1, 2, \ldots, n$) имеем вектор *описательных* (объясняющих) признаков $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})'$ (это, в частности, характеристики условий жизнедеятельности i-й обследованной семьи в примере В.1, значения параметров исследуемого технологического процесса, геофизических характеристик грунта или результаты обследований i-го пациента в задачах диагностики, геометрические или частотные характеристики распознаваемого образа в п. 3.3) и некоторую *информацию* Y_i *о* том *результирующем свойстве*, по которому производится классификация объектов (специфика социально-экономического поведения i-й семьи в примере В.1;

Таблица В.3

№ п/п	Тип задачи классификации	Варианты (примеры) конечных прикладных целей исследования для данного типа задачи классификации
1	*Комбинационные группировки* и их непрерывные обобщения	Составление частотных таблиц и графиков, характеризующих распределение статистически обследованных объектов по градациям или интервалам группирования характеризующих их признаков (см. п. 1 в примере В.1)
2	*Простая типологизация:* выявление «стратификационной структуры» множества статистически обследованных объектов, «нащупывание» и описание четко выраженных скоплений («сгустков», «кластеров», «образов», классов) этих объектов в анализируемом многомерном пространстве и построение правила отнесения каждого нового объекта к одному из выявленных классов	2.1. Классификация как необходимый предварительный этап исследования, когда *до* проведения основной статистической обработки множества анализируемых данных (построения регрессионных моделей, оценки параметров генеральной совокупности и т. д.) добиваются расслоения этого множества на однородные (в смысле проводимого затем статистического анализа) порции данных (см. примеры В.1 и В.2) 2.2. Выявление и описание расслоенной природы анализируемой совокупности статистически обследованных объектов с целью формирования плана выборочных (например, экономико-социологических) обследований этой совокупности (см. пример В.3) 2.3. Первый шаг в построении связных типологий (см. п. 3 таблицы и пример В.1)
3	*Связная неупорядоченная типологизация:* исследование зависимостей между не поддающимися упорядочению классификациями одного и того же множества объектов в разных признаковых пространствах, одно из которых построено на *результирующих* (поведенческих) признаках (отражающих специфику функ-	3.1. Прогноз экономико-социологических ситуаций или отдельных социально-экономических показателей, включая задачу выявления так называемых типообразующих признаков, в том числе латентных, т. е. непосредственно не наблюдаемых (см. пример В.1) 3.2. Диагностика в промышленности, технике, георазведке, медицине [94, 156, 83, 115]

Продолжение

№ п/п	Тип задачи классификации	Варианты (примеры) конечных прикладных целей исследования для данного типа задачи классификации
	ционирования объекта, его социально-экономическое поведение, состояние здоровья и т п), а другое — на *описательных* (отражающих условия функционирования и другие характеристики, от которых могут зависеть значения результирующих признаков)	3 3 Автоматическое (машинное) распознавание образов — зрительных, слуховых [59, 74, 83]
4	*Связная упорядоченная типологизация:* модификация связной неупорядоченной типологизации (см п 3 таблицы), обусловленная дополнительным допущением, что классы, получаемые в пространстве результирующих (поведенческих) признаков, поддаются экспертному упорядочению по некоторому сводному (как правило, латентному, непосредственно не наблюдаемому) свойству эффективности функционирования, качеству, степени прогрессивности (оптимальности) поведения и т. п	Построение и интерпретация единого (сводного) латентного признака-классификатора в виде функции от исходных описательных признаков классификация химических элементов по заряду их атомного ядра (периодическая система Д И Менделеева), классификация сельских поселений по уровню их социально-экономического развития [128], построение фактора общей одаренности в педагогике и психологии [9], построение сводного показателя эффективности функционирования предприятия [87]; построение интегральной характеристики уровня мастерства спортсменов в игровых видах спорта [6]; оценка тяжести заболевания пациента и т д
5	*Структурная типологизация* дополнение и развитие простой типологизации (см. п 2 таблицы) в направлении изучения и описания структуры взаимосвязей полученных классов, включая построение соответствующих иерархических систем (на классифицируемых элементах и на классах элементов), анализ роли и места каждого элемента и класса в общей структурной	5 1 Классификация задач многоцелевого комплекса (крупной программы, научного направления производственного комплекса и т. п.) 5 2 Классификация элементов и подсистем по их функциональному назначению (производств — в территориально-производственном комплексе, территориальных единиц — в народнохозяйственном разделении труда и потребления, элементов организационных структур и т. п.)

Продолжение

№ п/п	Тип задачи классификации	Варианты (примеры) конечных прикладных целей исследования для данного типа задачи классификации
	классификационной схеме При этом структурная классификационная схема определяется составляющими ее классами (подсистемами) и характеристиками (правилами) их взаимодействия	5.3 Классификация *лиц*, принимающих решение, по их ролн и близости позиций в понимании ситуации и способе решения задачи 5.4 Классификация исследуемых признаков и анализ структуры связей между ними Примеры по всем подпунктам данного пункта таблицы можно найти в [11, гл. 4; 111, гл 5—8, 34, 158]
6	*Классификация динамических траекторий развития систем* типологизация траекторий многомерных временных рядов $X(t) = (x^{(1)}(t), \ldots, x^{(p)}(t))'$, средн компонент $x^{(j)}(t)$ которых могут быть как количественные, так и качественные переменные	Задачи биологической систематики [111], анализа типов динамики семейной структуры [50], анализа типов динамики потребительского поведения семей [154] и др.

наличие или отсутствие сбоев в i-м анализируемом технологическом процессе, месторождений полезных ископаемых на i-м обследованном участке, заболевания у i-го обследуемого пациента в задачах диагностики; конкретный содержательный смысл распознаваемого зрительного или слухового образа в задачах п. 3.3). Разница между задачами типа 3.1 и задачами 3.2 и 3.3 заключается в том, что в задачах прогноза экономико-социологических ситуаций информация Y_i об исследуемом результирующем свойстве объекта *не является окончательной*, т. е. не задает однозначно, как это делается в задачах 3.2 и 3.3, образа (класса, типа), к которому относится этот объект. Эта информация в задачах типа 3.1 носит лишь *промежуточный характер* и представляется, как правило, в виде вектора результирующих показателей $Y_i = (y_i^{(1)}, \ldots, y_i^{(q)})'$. Поэтому в отличие от задач 3.2 и 3.3 (в которых уже «на входе» задачи имеем распределение анализируемых объектов-векторов X_i по классам, что и составляет так называемую «обучающую выборку») в задачах типа 3.1 нужно предварительно осуществить простую типологизацию множества объектов $\{O_i\}$ $(i = 1, \ldots, n)$ в пространстве результирующих показателей и лишь затем исполь-

зовать полученные в результате этой типологизации классы в качестве обучающих выборок для построения классифицирующего правила в пространстве описательных признаков $\Pi(X)$.

«На выходе» же всех задач типа 3.1 — 3.3 должны быть 1) набор наиболее информативных объясняющих переменных (так называемых *типообразующих признаков*) $z^{(1)}(X)$, $z^{(2)}(X)$, ..., $z^{(p')}(X)$, которые либо отбираются по определенному правилу из числа исходных описательных признаков $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, либо строятся в качестве некоторых их комбинаций; 2) правило отнесения (*дискриминантная функция, классификатор*) каждого нового объекта O^*, заданного значениями своих описательных признаков X^*, к одному из заданных (или выявленных в процессе предварительной простой типологизации) в пространстве $\Pi(Y)$ классов или образов. При этом типообразующие признаки $Z = (z^{(1)}(X), ..., z^{(p')}(X))'$ и искомое правило классификации должны быть подобраны таким образом, чтобы обеспечивать наивысшую (в определенном смысле) точность решения задачи отнесения объекта к одному из анализируемых классов по заданным значениям его описательных признаков X.

Из сформулированных выше конечных целей классификации видно, что тематику разбиения многомерных данных на однородные (в определенном смысле) группы подчас трудно отделить от *задач снижения размерности исследуемых данных*. Однако прикладные цели методов снижения размерности не исчерпываются сформулированной выше задачей перехода от исходного набора описательных признаков $x^{(1)}$, ..., $x^{(p)}$ к существенно более скромному (по численному составу) набору так называемых типообразующих признаков $z^{(1)}(X)$, ..., $z^{(p')}(X)$, которые являются наиболее характерными, наиболее определяющими с точки зрения полноты и точности разбиения исследуемых объектов на классы.

Выделим в качестве основных следующие *типовые прикладные задачи снижения размерности* анализируемого признакового пространства, обслуживаемые соответствующими разделами многомерного статистического анализа.

I. Отбор наиболее информативных показателей (включая выявление латентных факторов). Речь идет об отборе из исходного (априорного) множества признаков $X = (x^{(1)}, ..., x^{(p)})'$ или построении в качестве некоторых комбинаций исходных признаков относительно небольшого числа p' переменных $Z(X) = (z^{(1)}(X), ..., z^{(p')}(X))'$, которые обладали бы свойством наибольшей информативности в смысле, определенном, как правило, некоторым специально подобранным для каждого конкретного типа задач *критерием*

информативности $I_{p'}(Z)$. Так, например, если критерий $I_{p'}(Z)$ «настроен» на достижение максимальной точности регрессионного прогноза некоторого результирующего количественного показателя y по известным значениям предикторных переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, то речь идет *о наилучшем подборе наиболее существенных предикторов в модели регрессии* [11, § 8.7]. Если же критерий $I_{p'}(Z)$ устроен таким образом, что его оптимизация обеспечивает наивысшую точность решения задачи отнесения объекта к одному из классов по значениям X его описательных признаков, то речь идет *о построении системы типообразующих признаков в задаче классификации* (см. § 1.4, 2.5, 2.6, гл. 11) или *о выявлении и интерпретации некоторой сводной (латентной) характеристики изучаемого свойства* (см. гл. 15). Наконец, критерий $I_{p'}(Z)$ может быть нацелен на максимальную *автоинформативность* новой системы показателей Z, т. е. на максимально точное воспроизведение всех исходных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ по сравнительно небольшому числу вспомогательных переменных $z^{(1)}, \ldots, z^{(p')}$ ($p' \ll p$). В этом случае говорят *о наилучшем автопрогнозе* и обращаются к моделям и методам факторного анализа и его разновидностей (см. гл. 13 и 14).

II. Сжатие массивов обрабатываемой и хранимой информации. Этот тип задач тесно связан с предыдущим и, в частности, требует в качестве одного из основных приемов решения построения экономной системы вспомогательных признаков, обладающих наивысшей автоинформативностью, т. е. свойством наилучшего автопрогноза (см. выше). В действительности при решении достаточно серьезных задач сжатия больших массивов информации (подобные задачи весьма актуальны и в плане необходимости минимизации емкостей носителей, на которых хранится *архивная* информация, и в плане экономии памяти ЭВМ при обработке *текущей* информации) используется сочетание методов классификации и снижения размерности. Методы классификации позволяют подчас перейти от массива, содержащего информацию по всем n статистически обследованным объектам, к соответствующей информации только по k эталонным образцам ($k \ll n$), где в качестве эталонных образцов берутся специальным образом отобранные наиболее типичные представители классов, полученных в результате операции разбиения исходного множества объектов на однородные группы. Методы же снижения размерности позволяют заменить исходную систему показателей $X = (x^{(1)}, \ldots, x^{(p)})'$ набором вспомогательных (наиболее автоинформативных) переменных $Z(X) = (z^{(1)}(X), \ldots, z^{(p')}(X))'$.

Таким образом, размерность информационного массива понижается от $p \times n$ до $p' \times k$, т. е. во многие десятки раз, если учесть, что p' и k обычно на порядки меньше соответственно p и n.

III. Визуализация (наглядное представление) данных. Вернемся к примеру В.1. При проведении простой типологизации семей в «пространстве поведения» приходится иметь дело с множеством точек (семей) в 98-мерном пространстве[1]. А для формирования рабочих гипотез, исходных допущений о геометрической и вероятностной природе совокупности анализируемых данных $Y_1, \ldots, Y_n$ важно было бы суметь «подсмотреть», как эти данные точки располагаются в анализируемом пространстве $\Pi(Y)$. В частности, уже на предварительной стадии исследования хотелось бы знать, распадается ли исследуемая совокупность точек на четко выраженные сгустки в этом пространстве, каково примерное число этих сгустков и т.д.? Но максимальная размерность «фактически осязаемого» пространства, как известно, равна трем Поэтому, естественно, возникает проблема: нельзя ли спроецировать анализируемые многомерные данные из исходного пространства на прямую, на плоскость, в крайнем случае — в трехмерное пространство, но так, чтобы интересующие нас специфические особенности исследуемой совокупности (например, ее расслоенность на кластеры), если они присутствуют в исходном пространстве, сохранились бы и после проецирования. Следовательно, и здесь речь идет о снижении размерности анализируемого признакового пространства, но снижении, во-первых, подчиненном некоторым специальным критериям и, во-вторых, оговоренном условием, что размерность редуцированного пространства не должна превышать трех. Аппарат для решения подобных задач называется в книге «*целенаправленным проецированием*» многомерных данных и излагается в гл. 18—20.

IV. Построение условных координатных осей (многомерное шкалирование, латентно-структурный анализ). В данном типе задач снижение размерности понимается иначе, чем прежде. До сих пор речь шла о подчиненном некоторым специальным целям переходе от заданной координатной системы X (т. е. от исходных переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$) к новой координатной системе $Z(X)$, размерность которой p' существенно меньше размерности p и оси которой $0z^{(1)}, \ldots, 0z^{(p')}$ конструируются с помощью соответствующих преобра-

[1] Размерность вектора Y, т. е. число статей, по которым фиксируются удельные семейные расходы в бюджетных обследованиях, равна 98.

зований осей $0x^{(1)}, \ldots, 0x^{(n)}$. Теперь же рассматриваем ситуацию, когда *исходной содержательно заданной координатной системы не существует вовсе*, а подлежащие статистическому анализу данные представлены в виде $(B.1')$, т. е. в виде матрицы $\mathbf{A}$ попарных отношений a_{ij} $(i, j = 1, 2, \ldots, n)$ между объектами (см. (В.1')). Ставится задача: для заданной, сравнительно невысокой, размерности p' определить вспомогательные условные координатные оси $0z^{(1)}, \ldots, 0z^{(p')}$ и способ сопоставления каждому объекту O_i его координат $z_i^{(1)}, \ldots, z_i^{(p')}$ в этой системе таким образом, чтобы попарные отношения $\widehat{a_{ij}}$ (Z) (например, попарные взаимные расстояния) между объектами, вычисленные исходя из их содержательного смысла *на базе этих условных координат*, в определенном смысле минимально бы отличались *от заданных величин* a_{ij} $(i, j = 1, 2, \ldots, n)$. В определенных условиях (в первую очередь в задачах педагогики, психологии, построения различных рейтингов[1] и т. п.) построенные таким образом условные переменные поддаются содержательной интерпретации и могут тогда рассматриваться в качестве латентных характеристик определенных свойств анализируемых объектов (такого типа задачи называют часто задачами *латентно-структурного анализа*). Снижение размерности происходит здесь в том смысле, что от исходного массива информации размерности $n \times n$ переходим к матрице типа «объект — свойство» (см. (В.1)) размерности $p' \times n$, где $p' \ll n$. Аппарат для решения подобных задач состоит из методов так называемого *многомерного шкалирования* и представлен в гл. 16.

В.3. Типологизация математических постановок задач классификации и снижения размерности

Целесообразность и эффективность применения тех или иных методов классификации и снижения размерности так же, как их предметная осмысленность, обусловлены конкретизацией базовой математической модели, т. е. математической постановкой задачи. Определяющим моментом в выборе математической постановки задачи является ответ на вопрос, *на какой исходной информации строится модель*. При этом исходная информация складывается из двух час-

[1] В подобных ситуациях элементы матрицы (В 1') отражают обычно результаты попарных сравнений объектов O_i и O_j по анализируемому результирующему свойству и, следовательно, представляют одну из возможных форм обучающей информации (см. В.3).

тей: 1) *из априорных сведений об исследуемых классах*; 2) *из информации статистической, выборочной*, т. е. так называемых *обучающих или частично обучающих выборок* (точные определения см. в § 2.1 и 9.1).

Априорные сведения об исследуемых генеральных совокупностях относятся обычно к виду или некоторым общим свойствам закона распределения исследуемого случайного вектора X в соответствующем пространстве и получаются либо из теоретических, предметно-профессиональных соображений о природе исследуемого объекта, либо как результат предварительных исследований. Получение выборочной исходной информации в экономике и социологии, как правило, связано с организацией системы экспертных оценок[1] или с проведением специального предварительного этапа, посвященного решению задачи простой типологизации анализируемых объектов в пространстве результирующих показателей (см. выше пример В.1).

Классификация задач разбиения объектов на однородные группы (в зависимости от наличия априорной и предварительной выборочной информации) и соответствующее распределение описания аппарата решения этих задач по главам и параграфам данной книги представлены в табл. В.4.

Математическая модель, лежащая в основе построения того или иного метода снижения размерности, включает в себя обычно три основных компонента

1. **Форма задания исходной информации.** Речь идет об ответе на следующие вопросы: а) в каком виде (т. е. в виде (В.1), (В.1′) или еще каком-либо) задана описательная информация об объектах? б) имеется ли среди исходных статистических данных обучающая информация, т. е. какие-либо сведения об анализируемом результирующем свойстве? в) если обучающая информация присутствует в исходных статистических данных, то в какой именно форме она представлена? Это могут быть, в частности, в привязке к объекту O_i $(i = 1, 2, \ldots, n)$: значения «зависимой» количественной переменной («отклика») y_i в моделях регрес-

[1] Медико-биологические и физико-технические задачи имеют в этом смысле определенное преимущество: там обучающие выборки можно получить с помощью специально организованного контрольного экспериментального исследования. Специфика же социально-экономических исследований до последнего времени практически исключала возможность использования идей и методов контролируемого и планируемого эксперимента *До последнего времени* потому, что проводящиеся сейчас в стране работы по созданию средств машинного эксперимента в экономике [100] дают надежду на возможность осуществления подходов, основанных на активном эксперименте в экономике, уже в ближайшем будущем.

Таблица В.4

Априорные сведения о классах (генеральных совокупностях)	Предварительная выборочная информация		
	нет информации	есть частично обучающие выборки	есть обучающие выборки
Некоторые самые общие предположения о законе распределения исследуемого вектора: гладкость, сосредоточенность внутри ограниченной области и т. п.	Классификация без обучения кластер-анализ, таксономия, распознавание образов «без учителя», нерархические процедуры классификации (гл. 5, 7, 8, 10—12)	Методы кластер-анализа, дополненные основанным на частично обучающих выборках выбором начальных приближений числа и центров классов, их ковариационных матриц (§ 7.3, гл. 9)	Непараметрические методы дискриминантного анализа (§ 3.2)
Различаемые генеральные совокупности заданы в виде параметрического семейства законов распределения вероятностей (параметры неизвестны)	Интерпретация исследуемой генеральной совокупности как смеси нескольких генеральных совокупностей «Расщепление» этой смеси с помощью методов оценивания неизвестных параметров (гл. 6)	Методы расщепления смеси, дополненные оценками, полученными на частично обучающих выборок Модификация методов кластер-анализа (гл 9)	Параметрические методы дискриминантного анализа (гл. 2—4)
Различаемые генеральные совокупности заданы однозначным описанием соответствующих законов распределения	Классификация при полностью описанных классах различение статистических гипотез (гл 1)	Обучающие выборки не нужны	

сии; номер одиородного по анализируемому свойству класса, к которому отиосится объект O_i в задаче классификации; порядковый иомер (раиг) объекта O_i в ряду всех объектов, упорядочеиных по степени проявления рассматриваемого свойства, в задачах анализа предпочтеиий и построения упорядочеиных типологизаций; наконец, значения $Y_i = (y_i^{(1)}, \ldots, y_i^{(q)})'$ набора результирующих признаков, характеризующих анализируемое в классификациоиной задаче свойство (см. пример В.1)[1].

2. **Тип оптимизируемого критерия $I_{p'}(Z)$ информативности искомого набора признаков $Z = (z^{(1)}, \ldots, z^{(p')})'$.** Как уже отмечалось, критерий информативности может быть ориентирован на достижение разных целей.

Следует выделить целый класс *критериев автоинформативности*, т. е. критериев, оптимизация которых приводит к набору вспомогательных переменных $Z = (z^{(1)}, \ldots, z^{(p')})'$, позволяющих максимально точно воспроизводить (в том или ином смысле, в зависимости от конкретного вида критерия) информацию, содержащуюся в описательном массиве данных типа (В.1) или (В.1′). Если описательная информация представлена в виде матрицы «объект — свойство» (В.1), то речь идет о максимально точном восстановлении $p \times n$ значений исходных переменных $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}$ по значениям существенно меньшего числа $(p' \times n)$ вспомогательных переменных $z_i^{(1)}, \ldots, z_i^{(p')}$. Если же описательная информация представлена в виде матрицы попарных сравнений объектов (В.1′), то речь идет о максимально точном воспроизведении n^2 элементов этой матрицы a_{ij} $(i, j = 1, \ldots, n)$ по значениям существенно меньшего числа $(p' \times n)$ вспомогательных переменных $z_i^{(1)}, \ldots, z_i^{(p')}$ $(i = 1, \ldots, n)$.

Будем называть *критериями внешней информативности* (имеется в виду информативность, внешняя по отношению к информации, содержащейся в описательном массиве (В.1) или (В.1′)), такие критерии $I_{p'}(Z)$, которые нацелены на поиск экономных наборов вспомогательных переменных $Z(X) = (z^{(1)}(X), \ldots z^{(p')}(X))'$, обеспечивающих максимально точное воспроизведение (по значениям Z, а значит в конечном счете по значениям X) информа-

---

[1] Перечисленные варианты не исчерпывают всех возможных форм представления обучающей информации. Так, например, при надлежащей интерпретации элементов a_{ij} матрицы (В.1′) она может быть также отнесена к разновидностям обучающей информации (если, скажем, a_{ij} понимать как результат сравнения по анализируемому результирующему свойству объектов O_i и O_j).

ции, относящейся к *результирующему* признаку (варианты ее задания перечислены выше, в п.1)).

3. **Класс $\mathbf{L}(X)$ допустимых преобразований исходных признаков** X. Вспомогательные признаки $Z = (z^{(1)}, \ldots, z^{(p')})'$ в случае представления исходной описательной информации в форме матрицы «объект — свойство» (т. е. в виде (В.1)) конструируются в виде функций от X, т. е. $Z = Z(X)$. Как обычно в таких ситуациях, чтобы обеспечить содержательность и конструктивную реализуемость решения оптимизационной задачи

$$I_{p'}(Z(X)) \to \underset{Z(X)}{\text{extr}}^{1},$$

следует предварительно договориться об ограниченном классе допустимых решений $\mathbf{L}(X)$, в рамках которого эта оптимизационная задача будет решаться. Очевидно, от выбора $\mathbf{L}(X)$ будет существенно зависеть и получаемое решение $Z(X) = (z^{(1)}(X), \ldots, z^{(p')}(X))$ упомянутой оптимизационной задачи.

Итак, следуя предложенной выше логике, мы должны были бы произвести типологизацию задач снижения размерности по трем «входам» (или «срезам»): форме задания исходной информации, типу (смыслу) оптимизируемого критерия информативности и классу допустимых преобразований исходных переменных. Однако в предлагаемой ниже форме представления результатов типологизации задач снижения размерности (табл. В.5) эти принципы реализованы в упрощенном виде за счет следующих двух практических соображений: 1) подавляющее большинство методов снижения размерности базируется на *линейных* моделях, т. е. класс допустимых преобразований $\mathbf{L}(X)$ — это класс линейных (как правило, подходящим образом нормированных) преобразований исходных признаков $x^{(1)}, \ldots, x^{(p)}$ (в книге нелинейным преобразованиям посвящены лишь § 13.6 и 17.3); 2) спецификация формы задания исходной информации связана со спецификацией смысловой нацеленности критерия информативности, а поэтому их удобнее давать в общей графе.

Данная в табл. В.5 типологизация, как и всякая иная классификация, не претендует на исчерпывающую полноту. Заметим, что пункт 9 этой таблицы повторяет, по существу,

[1] Речь идет о нахождении (в виде функции от X) такого вектора $Z(X) = (z^{(1)}(X), \ldots, z^{(p')}(X))$, который обращает в максимум или минимум (в зависимости от конкретного содержательного смысла оптимизируемого критерия информативности) значение $I_{p'}(Z)$. Поэтому справа в данном соотношении записано extr («экстремум»).

Таблица В.5

№ п/п	Класс и смысловая нацеленность критерия информативности, форма задания исходной информации	Название соответствующих моделей и методов, главы и параграфы книги
1	АИ: максимизация содержащейся в $z^{(1)}, \ldots, z^{(p')}$ доли суммарной вариабельности исходных признаков $x^{(1)}, \ldots, x^{(p)}$. Описательная информация: в форме (В.1). Обучающая информация: нет	Метод главных компонент, гл. 13
2	АИ: максимизация точности воспроизведения корреляционных связей между исходными признаками по их аппроксимациям с помощью вспомогательных переменных $z^{(1)}, \ldots, z^{(p')}$. Описательная информация: в форме (В.1) Обучающая информация: нет	Модели и методы факторного анализа, гл. 14
3	АИ: разбиение исходных признаков на группы высококоррелированных (внутри группы) переменных и отбор от каждой группы фактора, имеющего максимальную интегральную характеристику корреляционных связей со всеми признаками данной группы Описательная информация: в форме (В.1) Обучающая информация: нет	Метод экстремальной группировки параметров, п. 14.2.1
4	АИ: приписывание каждому объекту O_t значений условных координат $\left(z_t^{(1)}, \ldots, z_t^{(p')}\right)$ таким образом, чтобы по ним максимально точно восстанавливалась заданная структура попарных описательных отношений между объектами. Описательная информация: в форме (В.1'). Обучающая информация: нет	Многомерное шкалирование, гл. 16

Продолжение

№ п/п	Класс и смысловая нацеленность критерия информативности, форма задания исходной информации	Название соответствующих моделей и методов, главы и параграфы книги
5	АИ: максимальное сохранение заданных описательным массивом (B.1) анализируемых структурно-геометрических и вероятностных свойств после его проецирования в пространство меньшей размерности (в пространство, натянутое на $z^{(1)}, \ldots, z^{(p')}$, $p' < p$). Описательная информация: в форме (B.1). Обучающая информация: нет	Методы целенаправленного проецирования и отбор типообразующих признаков в кластер-анализе, гл. 11, 18—21
6	ВИ: минимизация ошибки прогноза (восстановления) значения результирующей количественной переменной по значениям описательных переменных (предикторов). Описательная информация: в форме (B.1). Обучающая информация: в форме зарегистрированных на объектах $O_1, \ldots, O_n$ значений соответственно $y_1, \ldots, y_n$ результирующего количественного показателя y	Отбор существенных предикторов в регрессионном анализе, см. [11, § 8.7]
7	ВИ: минимизация вероятностей ошибочного отнесения объекта к одному из заданных классов по значениям его описательных переменных Описательная информация: в форме (B.1). Обучающая информация: для каждого описанного с помощью (B.1) объекта указан номер класса, к которому он относится	Отбор типообразующих признаков в дискриминантном анализе, § 1.4, 2.5, 2.6
8	ВИ: максимизация точности воспроизведения (по значениям вспомогательных признаков) заданных в «обучении» отношений объектов по анализируемому результирующему свойству. Описательная информация: в форме (B.1).	Методы латентно-структурного анализа, в том числе построение некоторой сводной латентной характе-

Продолжение

№ п/п	Класс и смысловая нацеленность критерия информативности, форма задания исходной информации	Название соответствующих моделей и методов, главы и параграфы книги
	Обучающая информация, в форме попарных сравнений или упорядочений объектов по анализируемому результирующему свойству (см. сноску к с. 36 о возможности использования формы (В.1′) для представления обучающей информации)	ристики изучаемого результирующего свойства, гл. 15
9	ВИ: максимизация точности воспроизведения (по значениям условных вспомогательных переменных) заданных в «обучающей информации» попарных отношений объектов по анализируемому результирующему свойству. Описательная информация: нет. Обучающая информация: в форме (В.1′) (см. сноску к с. 36)	Многомерное шкалирование как средство латентно-структурного анализа, гл. 16

Примечание. АИ — автоинформативность, ВИ — внешняя информативность.

пункт 4, они отличаются только интерпретацией исходных данных вида (В.1′) и соответственно конечными прикладными целями исследования.

В.4. Основные этапы в решении задач классификации и снижения размерности

Целью данного параграфа является конкретизация сформулированных в [12, п. 1.1.3] общих рекомендаций по методике проведения всякого статистического анализа данных. В этой конкретизации будем опираться на описанную выше специфику задач классификации и снижения размерности, и в частности на имеющуюся теперь возможность выбора подходящего типа практической задачи и соответствующих ему конечных прикладных целей исследования (см. § В.2), а также подбора необходимого математического инструментария (см. § В.3).

Представим весь процесс решения задач классификации и снижения размерности в виде следующей схемы (рис. В.1) и прокомментируем ее.

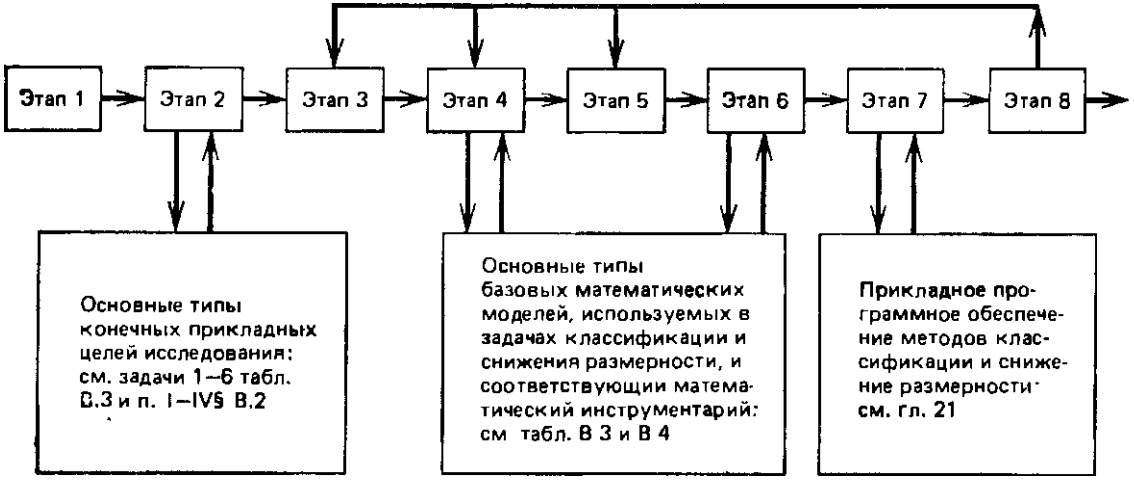


Рис. В.1. Схема поэтапного процесса решения задач классификации и снижения размерности

Этап 1 (установочный). На этом этапе главную роль играет «заказчик», т. е. специалист той предметной области (экономики, социологии, геологии, медицины и т. д.), к которой относится решаемая задача. *На предметно-содержательном уровне* должна быть сформулирована постановка задачи, включающая в себя характер научных или практических выводов, которые требуется получить «на выходе» задачи (диагностический, прогнозный, аналитический и т. п.), описание предмета исследования, объектов статистического обследования, выделяемые для решения задачи ресурсы (время, трудозатраты) и т. д.

Этап 2 (постановочный). На этом этапе необходимо тесное сотрудничество «заказчика» и «инструментальщика», т. е. специалиста по прикладной статистике. Отправляясь от выработанной на этапе 1 предметно-содержательной *установки* на главные цели исследования, они должны сформулировать эти цели в терминах основных типов прикладных задач, рассматриваемых в теории статистических методов классификации и снижения размерности (см. § В 2). Необходимым условием успешной реализации этого этапа (и соответственно всего последующего статистического анализа) является полное взаимопонимание и согласие «заказчика» и «инструментальщика» в принятом решении (что достигается далеко не просто).

Этап 3 (информационный). Состоит в выработке и реализации плана сбора исходной статистической информации (если ее не представил «заказчик» уже на этапе 1), в подробной аттестации этой информации (объяснение способа сбора, формы представления и т. п.), в вводе исходных данных в ЭВМ, их выверке и редактировании.

Этап 4 (априорный математико-постановочный). На базе выводов и информации, полученных в результате реализации этапов 1—3, требуется осуществить предварительный (априорный, т. е. до проведения каких-либо расчетов) выбор базовых математических моделей, которые целесообразно использовать в математической постановке данной конкретной задачи (см. § В 3). При этом факторами, от которых решающим образом зависит выбор, являются, как уже знаем, характер конечных прикладных целей исследования, природа и форма исходных статистических данных.

Этап 5 (разведочный анализ). Этот этап составляют всевозможные методы предварительной статистической обработки, «прощупывания» исходных данных с целью выявления специфики их вероятностной и геометрической природы ([12, гл. 10 и 11], а также гл. 18—21 данной книги).

«На выходе» этапа должны быть уточненные сведения о физическом механизме генерирования наших исходных данных, а значит, о базовой математической модели этого механизма.

Этап 6 (*апостериорный математико-постановочный*). На этом этапе уточняется математическая постановка решаемой задачи с учетом выводов, полученных на предыдущем этапе.

Этап 7 (*вычислительный*). Производится вычислительная реализация намеченного использования выбранного на предыдущем этапе математического инструментария в решении задачи. При этом желательно воспользоваться типовым программным обеспечением (см. гл 21).

Этап 8 (*итоговый*). Анализируются и интерпретируются результаты проведенной статистической обработки (классы, факторы и т.п.). В зависимости от результатов этого анализа (достигнуты ли все намеченные на этапе 2 прикладные цели исследования, насколько естественно интерпретируемы полученные результаты, степень их достоверности и т. п.) либо формулируются окончательные научные или прикладные выводы, либо даются уточнения и дополнения к заданию и возвращаются к одному из предыдущих этапов обычно к этапу 3, 4 или 5).

ВЫВОДЫ

1. Разделение рассматриваемой совокупности объектов или явлений на однородные (в определенном смысле) группы называется *классификацией*. При этом термин «классификация» используют, в зависимости от контекста, для обозначения как *самого процесса* разделения, так и *его результата*. Это понятие тесно связано с такими терминами, как группировка, типологизация, систематизация, дискриминация, кластеризация, и является одним из основополагающих в практической и научной деятельности человека.

2. Переход от характеризующего состояние или функционирование некоторой совокупности объектов исходного массива данных к существенно *более лаконичному набору показателей*, отобранных из числа исходных или построенных с помощью некоторого их преобразования таким образом, чтобы минимизировать связанные с этим потери в информации (содержавшейся в исходном массиве данных относительно рассматриваемой совокупности объектов), составляет сущность *процесса снижения размерности*.

Этот процесс использует, в частности, логику и приемы классификации, и сам в свою очередь используется в классификационных процедурах.

3. В ситуациях, когда каждый из исследуемых объектов или явлений характеризуется большим числом разнотипных и стохастически взаимосвязанных параметров, и исследователь имеет возможность получить, или уже получил, результаты статистического обследования по этим параметрам целой совокупности таких объектов или явлений, для решения задач классификации и снижения размерности следует привлекать специальный математический инструментарий *многомерного статистического анализа*: дискриминантный и кластер-анализ, методы расщепления смесей распределений, методы иерархической классификации, многомерное шкалирование, главные компоненты, факторный анализ, целенаправленное проецирование многомерных данных и т. п. Практическая реализация этих методов требует весьма сложных и трудоемких расчетов и стала возможной приблизительно лишь к середине нашего столетия, когда была создана необходимая вычислительная база.

4. К числу основных методологических принципов, которые лежат в основе большинства конструкций многомерного статистического анализа, следует отнести: а) необходимость учета *эффекта существенной многомерности анализируемых данных* (используемые в конструкциях характеристики должны учитывать структуру и характер статистических взаимосвязей исследуемых признаков); б) *возможность лаконичного объяснения природы анализируемых многомерных структур* (допущение, в соответствии с которым существует сравнительно небольшое число определяющих, подчас латентных, т. е. непосредственно не наблюдаемых, факторов, с помощью которых могут быть достаточно точно описаны все наблюдаемые исходные данные, структура и характер связей между ними); в) *максимальное использование «обучения»* в настройке математических моделей классификации и снижения размерности (под «обучением» понимается та часть исходных данных, в которой представлены «статистические фотографии» соотношений «входов» и «выходов» анализируемой системы); г) *возможность оптимизационной формулировки задач многомерного статистического анализа* (в том числе задач классификации и снижения размерности), т. е. нахождение наилучшей процедуры статистической обработки данных с помощью оптимизации некоторого экзогенно заданного критерия качества метода.

Первые два принципа относятся к природе обрабатываемых данных, а следующие два — к логике построения соответствующих аппаратных средств.

5. Среди типов прикладных задач (конечных прикладных целей) классификации следует выделить: 1) *комбинационные группировки* и их непрерывные обобщения — разбиение совокупности на интервалы (области) группирования; 2) *простая типологизация*: выявление естественного расслоения анализируемых данных (объектов) на четко выраженные «сгустки» (кластеры), лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные друг от друга части; 3) *связная неупорядоченная типологизация*: использование реализованной в пространстве *результирующих* показателей простой типологизации в качестве обучающих выборок при классификации той же совокупности объектов в пространстве *описательных* признаков; 4) *связная упорядоченная типологизация*, которая отличается от связной неупорядоченной возможностью *экспертного упорядочения* классов, полученных в пространстве результирующих показателей, и использованием этого упорядочения для построения сводного латентного результирующего показателя как функции от описательных переменных; 5) *структурная типологизация* дает на «выходе» задачи дополнительно к описанию классов еще и описание существующих между ними и их элементами структурных (в том числе иерархических) связей; 6) *типологизация динамических траекторий системы*: в качестве классифицируемых объектов выступают характеристики динамики исследуемых систем, например дискретные или непрерывные временные ряды или траектории систем, которые в каждый момент времени могут находиться в одном из заданных состояний.

6. Основные типы прикладных задач снижения размерности: 1) *отбор наиболее информативной системы показателей* (в задачах регрессии, классификации и т.п.); 2) *сжатие больших массивов информации*; 3) *визуализация* (наглядное представление многомерных данных); 4) *построение условного координатного пространства*, в терминах переменных которого в некотором смысле наилучшим образом описываются и интерпретируются анализируемые свойства объектов рассматриваемой совокупности.

7. При выборе подходящего математического инструментария для решения конкретной задачи *классификации* следует исходить из согласованного с «заказчиком» типа конечных прикладных целей исследования и характера априорной и выборочной информации (см. табл. В.4); при определении

математической модели, лежащей в основе выбора метода решения задачи *снижения размерности*, следует идти от типа прикладной задачи (см. предыдущий пункт выводов) к характеристике состава и формы исходных данных, а затем — к смысловой нацеленности и конкретному виду подходящего критерия информативности (см. табл. В.5).

8. Вся процедура статистического исследования, нацеленного на решение задачи классификации или снижения размерности, может быть условно разбита на восемь этапов (см. рис. В.1): 1) *установочный* (предметно-содержательное определение целей исследования); 2) *постановочный* (определение типа прикладной задачи в терминах теории классификации и снижения размерности); 3) *информационный* (составление плана сбора исходной информации и его реализация, если ее не было уже на этапе 1, затем предварительный анализ исходной информации, ее ввод в ЭВМ, сверка, редактирование); 4) *априорный математико-постановочный* (осуществляемый до каких бы то ни было расчетов выбор базовой математической модели механизма генерации исходных данных); 5) *разведочный* (специальные методы статистической обработки исходных данных, например целенаправленное проецирование, нацеленные на выявление их вероятностной и геометрической природы); 6) *апостериорный математико-постановочный* (уточнение выбора базовой математической модели с учетом результатов предыдущего этапа); 7) *вычислительный* (реализация на ЭВМ уточненного на предыдущем этапе плана математико-статистического анализа данных); 8) *итоговый* (подведение итогов исследования, формулировка научных или практических выводов).

Раздел I. ОТНЕСЕНИЕ К ОДНОМУ ИЗ НЕСКОЛЬКИХ КЛАССОВ, ЗАДАННЫХ ПРЕДПОЛОЖЕНИЯМИ И ОБУЧАЮЩИМИ ВЫБОРКАМИ

Глава 1. КЛАССИФИКАЦИЯ В СЛУЧАЕ, КОГДА РАСПРЕДЕЛЕНИЯ КЛАССОВ ОПРЕДЕЛЕНЫ ПОЛНОСТЬЮ

1.1. Два класса, заданные функциями распределения

1.1.1. Критерий отношения правдоподобия как правило классификации. В настоящей главе наблюдение $X' = (x^{(1)}, \ldots, x^{(p)})$ всегда является упорядоченным набором из p признаков-координат. Событие, что наблюдение извлечено из j-го класса, а также соответствующая гипотеза обозначаются H_j; распределение вектора X, принадлежащего j-му классу ($j = 1, \ldots, k$), обозначается $F_j(\ldots) \equiv F(\ldots \mid H_j)$, плотности вероятностей (вероятности) — соответственно $f_j(\ldots) \equiv f(\ldots \mid H_j)$.

Задача построения классификационных правил рассматривается при двух способах задания распределений X в классах: аналитическом, когда непосредственно задаются F_j с помощью подходящей математической формулы, и выборочном, когда распределения в классах задаются с помощью указания соответствующих генеральных совокупностей. Сюда в принципе можно было бы отнести и случаи дискриминантного анализа с выборками настолько большого объема, что выборочными флуктуациями используемых статистик можно пренебречь (§ 1.3).

Задача отнесения наблюдения X в один из двух ранее известных классов $j = 1, 2$ тесно связана с классической статистической задачей проверки простой гипотезы против простой альтернативы [11, § 9.3]. Например, гипотезы H_1: $X \in F_1$ против гипотезы H_2: $X \in F_2$. Известно (лемма Неймана — Пирсона), что в достаточно широком классе ситуаций [88] среди всех возможных критериев с *ошибкой первого рода* α наиболее мощным, т. е. имеющим наименьшую *ошибку второго рода* β, является критерий отношения правдоподобия, основанный на статистике

$$\gamma(X) = \frac{L(X \mid H_2)}{L(X \mid H_1)} = \frac{f_2(X)}{f_1(X)}, \tag{1.1}$$

где L — функция правдоподобия [11, с. 269].

При этом при $\gamma(X) \leqslant c_\alpha$ принимается гипотеза H_1, а при $\gamma(X) > c_\alpha$ принимается гипотеза H_2. Таким образом, R — пространство возможных значений X — с помощью $\gamma(X)$ разбивается на две непересекающиеся области: $K_1 = \{X : \gamma(X) \leqslant c_\alpha, \quad X \in R\}$ — область принятия H_1 и $K_2 = \{X : \gamma(X) > c_\alpha, \quad X \in R\}$ — область принятия H_2, или, как принято говорить в статистической теории проверки гипотез, *критическую область* для гипотезы H_1.

Пусть $\pi_j = P\{H_j\}$ означает априорные вероятности гипотез. Правило классификации

$$\frac{\pi_2 f_2(X)}{\pi_1 f_1(X)} \gtrless 1 \Rightarrow \begin{cases} H_2 \\ H_1 \end{cases} \tag{1.2}$$

называется *байесовским*. Очевидно, оно является частным случаем критерия отношения правдоподобия.

Рассмотрим произвольный критерий проверки гипотезы H_1 с критической областью (областью принятия гипотезы H_2) — K. Тогда по формуле полной вероятности [11, формула (4.14)] вероятность принять ошибочное решение

$$P\{\text{ошибка}\} = \pi_1 P\{\text{ошибка} \mid H_1\} + \pi_2 P\{\text{ошибка} \mid H_2\} =$$

$$= \pi_1 \int_K f_1(X)\, dX + \pi_2 \int_{\overline{K}} f_2(X)\, dX = \pi_1 \int_K f_1(X)\, dX +$$

$$+ \pi_2 \left(1 - \int_K f_2(X)\, dX\right) = \pi_2 + \int_K (\pi_1 f_1 - \pi_2 f_2)\, dX. \tag{1.3}$$

Интеграл в правой части (1.3) принимает наименьшее значение в случае, когда область K состоит из всех точек, где подынтегральная функция отрицательна, т. е. $\pi_1 f_1(X) < \pi_2 f_2(X)$, но это и есть определение байесовского классификатора. Таким образом, *байесовский классификатор минимизирует вероятность принятия ошибочного решения.*

Как будет видно из последующего материала, большинство используемых на практике алгоритмов классификации строится исходя из формулы (1.1). При этом либо оцениваются неизвестные параметры Θ предполагаемых теоретических распределений и вместо $\boldsymbol{\Theta}$ в плотности подставляются оценки $\widehat{\boldsymbol{\Theta}}$ и далее вычисляется оценка $\gamma(X)$ как

$$\widehat{\gamma}(X) = \frac{f_2(X, \widehat{\boldsymbol{\Theta}}_2)}{f_1(X, \widehat{\boldsymbol{\Theta}}_1)}. \tag{1.4}$$

Это так называемые *параметрические методы* построения алгоритмов классификации. Либо для данной точки X сразу, минуя оценку параметров Θ, строится оценка отношения $f_2(X)/f_1(X)$. Это так называемые *непараметрические методы.*

Введем несколько моделей, используемых в теоретических исследованиях задачи классификации, и применим к ним критерий отношения правдоподобия для получения соответствующих критических областей. При этом удобно вместо $\gamma(X)$ использовать $h(X) = \ln \gamma(X)$.

1.1.2. Основные математические модели. Ниже, там, где это не вызывает недоразумений, для случайной величины и ее конкретного значения будет использоваться одно и то же обозначение X. Это позволит сделать формулы более обозримыми. При этом запись $f(X)$ в случае непрерывного распределения X будет означать плотность распределения случайной величины X в точке X, а в случае дискретного распределения X — соответственно вероятность того, что случайная величина X примет конкретное значение X. Рассмотрим четыре основные модели.

Модель двух дискретных распределений с независимыми координатами. В этом случае для $j = 1, 2$

$$L(X \mid H_j) = \prod_{k=1}^{p} P(x^{(k)} \mid H_j) = \prod_{k=1}^{p} f_j(x^{(k)}), \tag{1.5}$$

и области принятия гипотез K_j $(j = 1, 2)$ имеют вид

$$h(X) = \sum_{k=1}^{p} a_k(x^{(k)}) \gtrless c, \tag{1.6}$$

где c — некоторая постоянная и

$$a_k(x^{(k)}) = \ln(f_2(x^{(k)})/f_1(x^{(k)})). \tag{1.7}$$

Естественно трактовать $a_k(x^{(k)})$ как балл в пользу H_2 против H_1, приписанный соответствующему значению k-й координаты. Алгоритмы вида (1.5) из-за их простоты и наглядности часто используют в практической работе, хотя служащая их основанием модель весьма искусственна. Чтобы уменьшить влияние на результаты классификации несоответствия модели данным, в формуле (1.5) берут не все координаты X, а только их подмножество $x^{(i_1)}, \ldots, x^{(i_k)}$ $(k < p)$, подбирая $x^{(i_l)}$ так, чтобы вместе взятые они оставались достаточно информативными в отношении различения H_1 и H_2 и зависимость между ними (при фиксации гипотезы H_j $(j = 1, 2)$ была небольшой. Кроме того, для уменьшения эффекта зависимости при определении баллов $a_k(x^{(k)})$ (оцифровке значений $x^{(k)}$) для зависимых координат отступают от формулы (1.7), подбирая $a_k(x^{(k)})$ так, чтобы оптимизировать выбранный показатель качества классификации среди всех правил вида (1.6).

Разность ожидаемых значений $a_k(x^{(k)})$ при H_2 и H_1:

$$J_k = \mathbf{E}(a_k(x^{(k)}) \mid H_2) - \mathbf{E}(a_k(x^{(k)}) \mid H_1) =$$

$$= \sum \ln \frac{f_2(x^{(k)})}{f_1(x^{(k)})} (f_2(x^{(k)}) - f_1(x^{(k)})), \tag{1.8}$$

где суммирование проводится по всем возможным значениям $x^{(k)}$, рассматривают в качестве параметра, характеризующего среднюю информативность k-й координаты в различении гипотез H_1 и H_2. Основание для этого обсуждается в п. 1.2.4.

Модель двух дискретных распределений с одной и той же древообразной структурой зависимостей координат (ДСЗ-распределений). Функция правдоподобия для ДСЗ-распределений имеет вид [12, § 4.2]

$$L(X \mid H_j) = \prod_{l=1}^{p} f_j(x^{(\alpha(l))} \mid x^{(k(\alpha(l)))}) =$$

$$= \prod_{l=1}^{p} f_j(x^{(l)} \mid x^{(k(l))}), \tag{1.9}$$

где $\alpha = (\alpha(1), \ldots, \alpha(p))$ — некоторая перестановка координат вектора X, $k(\alpha(i)) \in \{0, \alpha(1), \ldots, \alpha(i-1)\}$ и $k(\ldots) = 0$ соответствует фиктивной координате $x^{(0)} = 1$. Применение критерия отношения правдоподобия дает области K_j ($j = 1, 2$) вида

$$h(X) = \sum_{l=1}^{p} b_l(x^{(l)}, x^{(k(l))}) \lesseqgtr c, \tag{1.10}$$

где

$$b_l(u, v) = \ln f_{2, x^{(l)}}(u \mid x^{(k(l))} = v) -$$

$$- \ln f_{1, x^{(l)}}(u \mid x^{(k(l))} = v). \tag{1.11}$$

Если координаты X (при фиксированной гипотезе H_1 или H_2) независимы, то оцифровки (1.7) и (1.11) совпадают. В литературе встречаются указания на большую практическую эффективность правил классификации, основанных на формулах (1.10) и (1.11), по сравнению с классификацией с помощью формул (1.6) и (1.7) [127].

Модель двух нормальных распределений с общей ковариационной матрицей (модель Фишера). Теоретические распределения в этом случае суть $N(M_1, \Sigma)$ и $N(M_2, \Sigma)$, причем $|\Sigma| > 0$. Правило классификации, соответственно K_1 и K_2, определяется с помощью неравенств

$$h(X) = (X - (M_1 + M_2)/2)' \Sigma^{-1} (M_2 - M_1) \lesseqgtr c. \tag{1.12}$$

Особенность модели Фишера состоит в том, что это простейшая математическая модель, допускающая произвольную ковариационную матрицу координат Σ, лишь бы только она не была вырожденной. Необычайно просто выглядит в модели и граница между областями принятия гипотез H_1 и H_2. Это гиперплоскость в p-мерном пространстве, касательная в одной и той же точке к одной из линий постоянного уровня плотности $N(M_1, \Sigma)$ и одной из линий постоянного уровня плотности $N(M_2, \Sigma)$ (рис. 1.1).

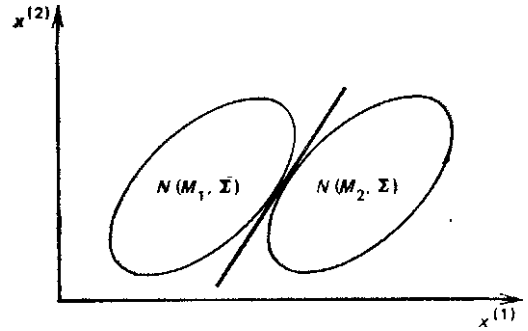


Рис. 1.1. Классификация плоскостью двух нормальных распределений с общей ковариационной матрицей

Модель двух нормальных распределений с разными ковариационными матрицами. Распределения в этом случае суть

$$N(M_j,\ \Sigma_j),\ |\Sigma_j| > 0,\ j = 1,\ 2. \tag{1.13}$$

Области K_1 и K_2 определяются выражением

$$h(X) = (X - M_1)'\ \Sigma_1^{-1}(X - M_1) - (X - M_2)'\ \Sigma_2^{-1}(X - M_2) + \\ + \ln(|\Sigma_1| / |\Sigma_2|) \lesseqgtr c. \tag{1.14}$$

Здесь $h(X)$ — полином второго порядка от координат X.

1.1.3. Классификация посредством задания границы критической области. Как показано в предыдущем пункте, для основных статистических моделей граница, разделяющая K_1 и K_2 — области принятия соответственно H_1 и H_2, выглядит достаточно просто. На практике в случаях, когда исходные распределения отличаются от базовых моделей, рассмотренных в предыдущем пункте, пренебрегают возможностью повышения эффективности классификации за счет точного следования критерию отношения правдоподобия (1.1) и ограничиваются областями принятия гипотез с гра-

ницами, принадлежащими какому-либо простому малопараметрическому семейству. При этом по-прежнему остается задача поиска критерия, наилучшего в заданном смысле (см. п. 1.1.4) среди допустимых (предположениями о границе) областей.

Классификация посредством линейной гиперплоскости. Рассмотрим модель (1.13) двух нормальных распределений с различными средними и ковариационными матрицами и попытаемся найти гиперплоскость $h(X) = V'X + v_0 = 0$ такую, чтобы критерий вида

$$h(X) \lesseqgtr 0 \Rightarrow \begin{cases} H_1 \\ H_2 \end{cases} \tag{1.15}$$

минимизировал ошибку классификации второго рода β при заданной ошибке классификации первого рода α [178]. Введем необходимые обозначения. Пусть для $j = 1, 2$

$$a_j = \mathbf{E}(h(X) \mid H_j) = V' \mathbf{E}(X \mid H_j) + v_0 = V' M_j + v_0, \tag{1.16}$$

$$\sigma_j^2 = \mathbf{E}\{(h(X) - a_j)^2 \mid H_j\} = V' \mathbf{E}\{(X - M_j)(X - M_j)' \mid H_j\} V = V' \Sigma_j V. \tag{1.17}$$

Поскольку линейная комбинация нормально распределенных случайных величин распределена нормально, из (1.15) — (1.17) следует, что

$$\alpha = P\{h(X) > 0 \mid H_1\} = 1 - \Phi(-a_1/\sigma_1), \tag{1.18}$$

$$\beta = P\{h(X) < 0 \mid H_2\} = \Phi(-a_2/\sigma_2), \tag{1.19}$$

где $\Phi(t) = (2\pi)^{-1/2} \int\limits_{-\infty}^{t} \exp\{-u^2/2\}\, du$.

Для отыскания V и v_0 воспользуемся методом множителей Лагранжа. Пусть $\psi = \Phi(-a_2/\sigma_2) + \lambda(1 - \Phi(-a_1/\sigma_1) - \alpha)$, тогда

$$\frac{\partial \psi}{\partial V} = \dot{\Phi}(-a_2/\sigma_2) \frac{\partial}{\partial V}(-a_2/\sigma_2) - \lambda \dot{\Phi}(-a_1/\sigma_1) \frac{\partial}{\partial V}(-a_1/\sigma_1) =$$

$$= \dot{\Phi}(-a_2/\sigma_2)(M_2 - a_2 \sigma_2^{-2} \Sigma_2 V)/\sigma_2 - \lambda \dot{\Phi}(-a_1/\sigma_1) \times$$

$$\times (M_1 - a_1 \sigma_1^{-2} \Sigma_1 V)/\sigma_1 = 0; \tag{1.20}$$

$$\frac{\partial \psi}{\partial v_0} = \dot{\Phi}(-a_2/\sigma_2)/\sigma_2 - \lambda \dot{\Phi}(-a_1/\sigma_1)/\sigma_1 = 0; \tag{1.21}$$

$$\frac{\partial \psi}{\partial \lambda} = 1 - \Phi(-a_1/\sigma_1) - \alpha = 0. \tag{1.22}$$

Исключив из уравнения (1.20) с помощью уравнения (1.21) множитель λ, получаем

$$M_2 - M_1 = [a_2 \sigma_2^{-2} \Sigma_2 - a_1 \sigma_1^{-2} \Sigma_1] V. \quad (1.23)$$

Предположим для простоты, что хотя бы одна из матриц Σ_j ($j = 1, 2$) положительно определена и что α и β меньше 0,5. Тогда, как нетрудно видеть, $a_1 < 0$, $a_2 > 0$, матрица, стоящая в квадратных скобках в правой части (1.23), положительно определена и имеет обратную. Воспользуемся последним обстоятельством для решения системы (1.20) — (1.22). Обозначим $b = a_2 \sigma_2^{-2} - a_1 \sigma_1^{-2}$, $s = -a_1 \sigma_1^{-2}/b$. В сделанных выше предположениях $b > 0$ и $0 < s < 1$. Из (1.23) следует, что

$$V = b^{-1} [s\Sigma_1 + (1-s)\Sigma_2]^{-1} (M_2 - M_1). \quad (1.24)$$

Далее, заменив a_j ($j = 1, 2$) по формулам (1.16) в определении s, получаем

$$v_0 = -(s\sigma_1^2 V' M_2 + (1-s)\sigma_2^2 V' M_1)/(s\sigma_1^2 + (1-s)\sigma_2^2). \quad (1.25)$$

Вычислительная процедура теперь может быть следующей:

1) для каждого $0 < s < 1$ при $b = 1$ вычисляется значение $V(s)$ по формуле (1.24) и далее последовательно по формулам (1.17), (1.25), (1.16), (1.18), (1.19) находятся $\sigma_j^2(s)$, $v_0(s)$, $a_j(s)$, $\alpha(s)$, $\beta(s)$;

2) на двумерной плоскости (u, v) строится график кривой $u = \alpha(s)$, $v = \beta(s)$ $(0 < s < 1)$;

3) пусть этот график пересекается с прямой $u = \alpha$ при $s = s_0$. Тогда искомый критерий

$$h(X) = V'(s_0) X + v_0(s_0) \quad \text{и} \quad \beta = \beta(s_0). \quad (1.26)$$

Достоинство этой процедуры состоит в том, что для настройки используется только один параметр s, а не $p + 1$ параметров, как при поиске решения напрямую в пространстве (V, v_0). Одновременное приведение к диагональному виду матриц Σ_1 и Σ_2 в начале работы дает дальнейшую экономию общего объема вычислений.

Кусочно-линейные классификаторы. Пусть пространство наблюдений R^p разбито на k взаимно непересекающихся подобластей R_i ($i = 1, \ldots, k$): $R_i \cap R_j = \varnothing$ для $i \neq j$ и $\bigcup R_i = R^p$; $h_i(X) = V_i' X + v_{i0}$, $i = 1, \ldots, k$, — уравнения линейных плоскостей. Классификатор вида

$$H(X) = h_t(X) \gtrless 0 \Rightarrow \begin{cases} H_2, \\ H_1, \end{cases} \text{ где } t = i : X \in R_i, \quad (1.27)$$

будем называть *кусочно-линейным* [44, с. 94—95].

Один из приемов приближенного малопараметрического описания многомерных распределений заключается в том, что их представляют в виде *конечной смеси однотипных нормальных законов*, отличающихся только параметрами сдвига

$$F(U) \approx \Sigma \omega_i N(U, A_i, \mathbf{I}_p) \quad (\Sigma \omega_i = 1) \tag{1.28}$$

или

$$F(U) \approx \Sigma \omega_i N(U, A_i, \Sigma) \quad (\Sigma \omega_i = 1). \tag{1.29}$$

При применении преобразования $Y = \Sigma^{-1/2} X$ (1.29) сводится к (1.28). В практической работе наиболее часто используется представление (1.28) [166, 168, 169], при этом векторы называют центрами или *эталонами*.

Рассмотрим задачу классификации распределений $F(U) = \Sigma \omega_{i1} N(U, A_i, \mathbf{I}_p)$ (гипотеза H_1) и $G(U) = \Sigma \omega_{j2} N(U, B_j, \mathbf{I}_p)$ (гипотеза H_2).

Оптимальный критерий согласно (1.1) должен задаваться с помощью

$$\gamma(X) = \frac{\Sigma \omega_{j2} \exp\{-(X - B_j)'(X - B_j)/2\}}{\Sigma \omega_{i1} \exp\{-(X - A_i)'(X - A_i)/2\}}. \tag{1.30}$$

На практике часто оставляют в суммах в числителе и знаменателе (1.30) по одному слагаемому, для которого соответствующий эталон наиболее близок к X, пренебрегают различиями в весах ω_i. При этом наблюдение X относится к той популяции, к наиближайшему эталону которой оно ближе. Полученный классификатор называется *кусочно-линейным классификатором по минимуму расстояния*. Разделяющая поверхность в этом случае является кусочно-линейной, состоящей из кусков гиперплоскостей. Вид разделяющей поверхности может быть разнообразным и зависит от взаимного расположения классифицируемых совокупностей (рис. 1.2).

Статистические вопросы, связанные с применением к моделям (1.28) описанного выше кусочно-линейного классификатора, исследовались в [168, 169].

1.1.4. Функция потерь. В предшествующих томах справочного издания [11, 12] уже неоднократно сталкивались с методическим приемом, когда для характеристики решения некоторой статистической задачи вводится подходящая функция потерь Q, а наилучшее (в смысле Q) решение определяется как решение, на котором при заданных ограничениях достигается минимум Q. Укажем основные функции потерь, ис-

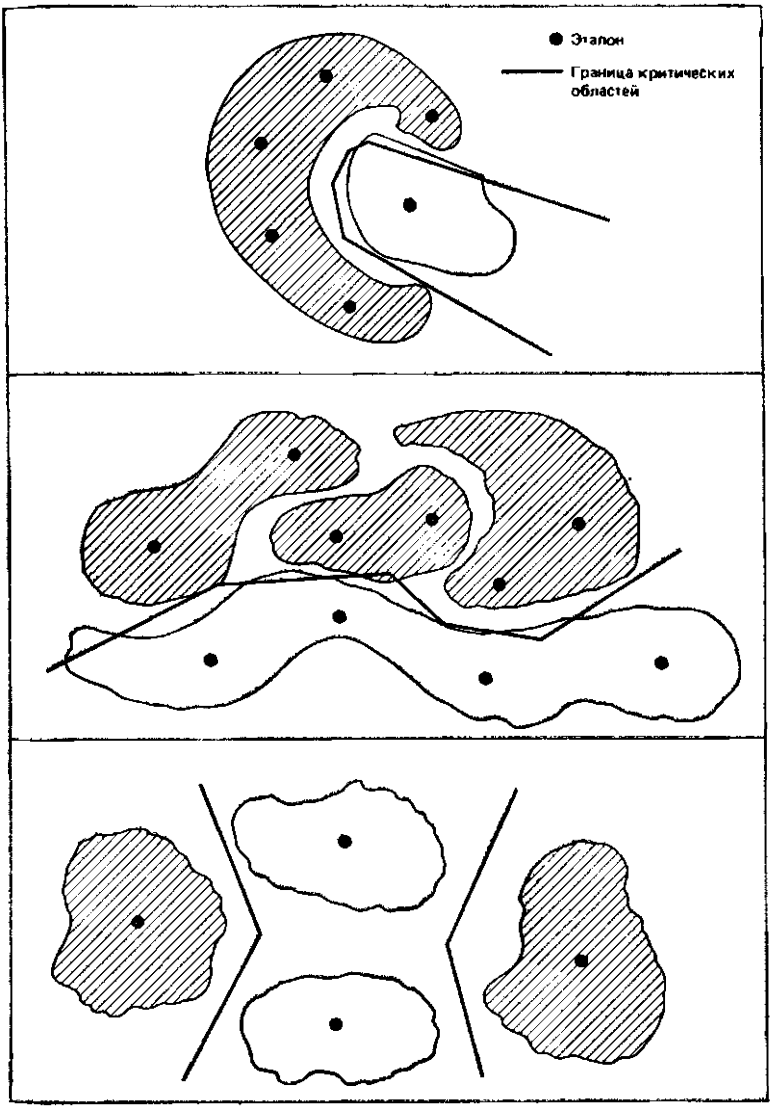


Рис. 1.2. Разделяющая поверхность кусочно-линейного классификатора по минимуму расстояния для трех случаев расположения классов

пользуемые в задаче классификации двух статистических распределений.

Вероятность ошибочной классификации (δ). Пусть, как в п. 1.1.1, π_i — априорная вероятность гипотезы H_j ($j = 1, 2$), тогда

$$\delta = P\{\text{ошибка}\} = \sum_{j=1}^{2} \pi_j P\{\text{ошибка} \mid H_j\} = \pi_1 \alpha + \pi_2 \beta. \quad (1.31)$$

Ввиду важности введенного понятия дадим его параллельное определение. Пусть $y = j$ в случае, когда верна гипотеза H_j ($j = 1, 2$), и $\widehat{y}(X)$ — решающая функция, которая тоже принимает два значения: $\widehat{y}(X) = j$, когда принимается гипотеза H_j, тогда δ может быть определена так же, как

$$\delta = \mathrm{E}(y - \widehat{y}(X))^2, \quad (1.31')$$

где математическое ожидание берется с учетом априорного распределения гипотез.

Частный случай формулы (1.31), получаемый при $\pi_j = 0{,}5$, дает полусумму ошибок $(\alpha + \beta)/2$. Как увидим в следующем параграфе, эта величина является удобной мерой разделения статистических совокупностей в случае модели Фишера.

На практике ошибки первого и второго рода не всегда эквивалентны. Так, например, при диспансеризации населения пропуск возможного заболевания более опасен, чем ложная тревога. Так возникает взвешенная ошибка классификации

$$Q = c_1 \pi_1 \alpha + c_2 \pi_2 \beta, \quad (1.32)$$

где c_j — штраф за ошибку, когда верна гипотеза H_j ($j = 1, 2$). Пусть y и $\widehat{y}(X)$ определены как выше и пусть

$$q_{c_1, c_2} = \begin{cases} 0 \\ c_1, \\ c^2 \end{cases} \text{ если } y - \widehat{y}(X) = \begin{cases} 0 \\ -1, \\ 1 \end{cases}$$

тогда по аналогии с (1.31′)

$$Q = \mathrm{E} q_{c_1, c_2}(X). \quad (1.32')$$

С точностью до постоянного множителя (1.32) эквивалентно (1.31), но с другим априорным распределением $\pi_i' = c_i \pi_i / \sum_{j=1}^{2} c_j \pi_j$.

На практике используются также функции потерь, зависящие не только от y и его оценки $\widehat{y}(X)$, но и от условной вероятности $P\{\widehat{y}=y|X\}$ [238]. Потому что одно дело — допустить ошибку там, где сомневаешься в ответе, другое — там, где уверен. Простейшая функция потерь, зависящая от условной вероятности, имеет вид:

$$q(y, \widehat{y}(X), h(X)) = \begin{cases} 1 & h(X) \geqslant a, \quad y=2, \\ 1 & h(X) \leqslant -a, \ y=1, \\ 1-\dfrac{h(X)+a}{\Delta}, & \text{если } -a \leqslant h(X) \leqslant \Delta - \\ & -a, \ y=1, \\ 1+\dfrac{h(X)-a}{\Delta} & a-\Delta \leqslant h(X) \leqslant a, \ y=2, \\ 0 \text{ в остальных случаях,} \end{cases}$$

где $0 < a < \Delta$ — некоторые постоянные, а $h(X) = \ln(\pi_2 P\{\widehat{y}=2|X\}/\pi_1 P\{\widehat{y}=1|X\})$.

1.1.5. Другие многомерные распределения. В теоретических и прикладных работах по классификации используется ряд многомерных распределений, в различных направлениях обобщающих многомерное нормальное распределение и его частные случаи. Укажем наиболее важные из них.

Эллипсоидальные распределения. Пусть $\xi \in R^p$ равномерно распределено на p-мерной сфере $C_p = \{X : X^TX=p\}$; η — неотрицательная случайная величина, не зависящая от ξ и имеющая строго возрастающую функцию распределения $F(u)$ такую, что $\int u^2 F(du) = 1$; $A \in R^p$; $\mathbf{B}$ — матрица невырожденного линейного преобразования R^p и $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}'$. Будем говорить, что случайная величина

$$\zeta = \eta \cdot \mathbf{B}\xi + A \tag{1.33}$$

имеет *эллипсоидальное распределение* $El(A, \mathbf{\Sigma}, F)$. Основанием для названия служит то, что, как и для нормального распределения, на концентрических эллипсоидах вида

$$(X-A)' \mathbf{\Sigma}^{-1} (X-A) = \text{const}, \tag{1.34}$$

где $A = \mathbf{E}\zeta$ и $\mathbf{\Sigma} = \mathbf{E}(\zeta - A)(\zeta - A)'$, плотность распределения ζ постоянна. В частном случае, когда $p\eta^2$ имеет χ^2-распределение с p степенями свободы, распределение ζ совпадает с нормальным $N(A, \mathbf{\Sigma})$. Это обстоятельство используется при статистическом моделировании случайных величин ξ. Так, если $\zeta \in N(0, \mathbf{I}_p)$, то $\xi = \sqrt{p}\,\zeta/(\zeta'\zeta)^{1/2}$ равномерно распределено на C_p

В модели независимых выборок из $El(M_1, \Sigma, F_1)$ и $El(M_2, \Sigma, F_2)$ при дополнительном предположении существования плотностей $f_j(u) = dF_j(u)/du$ отношение правдоподобия имеет вид:

$$\gamma(X) = t_2^{-p} f_2(t_2)/(t_1^{-p} f_1(t_1)),$$

где $t_j^2 = (X - M_j)' \Sigma^{-1} (X - M_j)$, $j = 1, 2$.

Откуда при $f_2(u) = f_1(u) = f(u)$ в случае, когда $u_2^{-p} \times \times f(u_2)/(u_1^{-p} f(u_1))$ — монотонная функция разности $u_2^2 - u_1^2$, общий вид классификатора максимального правдоподобия такой же, что и в (1.12). Сохраняется также и способ нахождения наилучшей разделяющей плоскости в модели независимых выборок из $El(M_1, \Sigma_1, F_1)$ и $El(M_2, \Sigma_2, F_2)$, $M_1 \neq M_2$, $\Sigma_1 \neq \Sigma_2$ (см. п. 1.1.3 и [25]).

Распределения, трансформируемые к нормальному. Пусть координаты вектора $\zeta = (\zeta^{(1)}, \ldots, \zeta^{(p)})'$ имеют непрерывные одномерные функции распределения $F_j(u) = P\{\zeta^{(j)} < u\}$ с плотностями соответственно $f_j(u) = dF_j(u)/du$ $(j = 1, \ldots, p)$. Будем говорить, что ζ имеет *трансформируемое к нормальному* (короче, *T-нормальное*) *распределение* $NT(X, \Sigma, F)$, где Σ — $(p \times p)$-неотрицательно определенная матрица, а $F'(X) = (F_1(x^{(1)}), \ldots, F_p(x^{(p)}))$ — вектор-функция одномерных распределений X, если

$$\xi = \Phi^{-1} F(\zeta) = \begin{pmatrix} \Phi^{-1}(F_1(\zeta^{(1)})) \\ \ldots\ldots\ldots \\ \Phi^{-1}(F_p(\zeta^{(p)})) \end{pmatrix} \in N(0_p, \Sigma),$$

где Φ^{-1} — функция, обратная $\Phi(u) = (2\pi)^{-1/2} \int_{-\infty}^{u} \exp\{-v^2/2\}\, dv$. Введем p одномерных функций $t_j(u) = F_j^{-1}(\Phi(u))$, тогда T-нормальное распределение можно также определить как распределение вектора $\zeta = (t_1(\xi^{(1)}), \ldots, t_p(\xi^{(p)}))'$, где $\xi = (\xi^{(1)}, \ldots, \xi^{(p)}) \in N(0, \Sigma)$.

Обозначим плотность T-нормального распределения $\varphi T(X, \Sigma, F)$. Предположим, что $|\Sigma| \neq 0$, и пусть $f^{(p)}(X) = \prod_{i=1}^{p} f_i(x^{(i)})$ и $q(X, \Sigma, F) = |\Sigma|^{-1/2} \exp\{-(\Phi^{-1}(F(X)))' \times \times (\Sigma^{-1} - I_p)(\Phi^{-1}(F(X)))/2\}$, тогда

$$\varphi T(X, \Sigma, F) = q(X, \Sigma, F) \cdot f^{(p)}(X). \qquad (1.35)$$

Пусть $\xi_1, \ldots, \xi_n$ — независимая выборка объема n из $NT(\Sigma, F)$, $\xi_k^{(i)}$ — i-я координата k-го наблюдения, $r_i(k)$ —

ее ранг в вариационном ряду i-х координат $\xi^{(i)}(1) < \ldots < \xi^{(i)}(j) < \ldots < \xi^{(i)}(n)$.

$\mathbf{R}(n) = \| r_i(k) \|$ — $(p \times n)$-матрица рангов и

$\Xi(n) = \| \xi^{(i)}(j) \|$ — $(p \times n)$-матрица вариационных рядов.

Замечательная особенность T-нормальных распределений заключается в том, что для оценки $\mathbf{F}$ надо использовать только матрицу $\Xi(n)$, а для оценки Σ — только матрицу $\mathbf{R}(n)$ [194].

Сформулированные выше модели выборок из нормальных распределений обобщаются на случай T-нормальных распределений. Так, аналог модели Фишера (см. п. 1.1.2) формулируется: даны две независимых выборки из $NT(\Sigma, \mathbf{F}_1)$ и $NT(\Sigma, \mathbf{F}_2)$, при этом известно, что для всех X

$$\Phi^{-1}(\mathbf{F}_1(X)) = \Phi^{-1}(\mathbf{F}_2(X)) + L,$$

где $L = (l^{(1)}, \ldots, l^{(p)})'$ — некоторый ненулевой вектор, и матрица Σ положительно определена.

Распределения с простой структурой связей между признаками. С простейшей моделью дискретных распределений с признаками, имеющими древообразную структуру зависимостей, познакомились в п. 1.1.2. Эта модель, естественно, может быть усилена предположением, что признаки имеют $R(k)$-распределение [12, § 4.4]. Однако без дополнительных предположений общий вид $\gamma(X)$ для $k > 1$ слишком сложен. Вместе с тем предположение о $R(k)$-зависимости признаков для нормальных распределений позволяет заметно уменьшить число параметров, от которых зависит ковариационная матрица, и это дает существенный выигрыш в ряде задач (см. пп. 1.4.1, 2.3.1 и 2.3.3).

Другое обобщение моделей с независимыми признаками — это параметрические модели, в которых вектор параметров Θ и вектор наблюдений X могут быть так разбиты на k взаимно непересекающихся подмножеств $\Theta' = \{\Theta^{(1)'}, \ldots, \Theta^{(k)'}\}$ и $X' = \{X^{(1)'}, \ldots, X^{(k)'}\}$, что плотность

$$f(X, \Theta) = \prod_{i=1}^{k} f_i(X^{(i)}, \Theta^{(i)}). \qquad (1.36)$$

Распределения, удовлетворяющие (1.36), будем называть *распределениями с независимыми блоками*. Они широко используются в теоретических исследованиях (см. пп. 2.3.2 и 2.5.3).

1.2. Характеристики качества классификации

Как уже выше сказано, с математической точки зрения задача классификации наблюдения X в одно из двух известных распределений F_j ($j = 1, 2$) сводится к проверке простой гипотезы «X принадлежит F_1» (или, короче, «$X \in F_1$») против простой альтернативы «X принадлежит F_2». Известно [11, § 9.2—9.4], что качество решения в этом случае описывается ошибками первого и второго рода. Однако ввиду высокой содержательной важности рассматриваемой задачи на практике используются более сложные формы заключений, такие, например, как трехградационное решение «$X \in F_1$», «отказ от классификаций», «$X \in F_2$» или указание условной вероятности $P\{X \in F_1 \mid X\}$. Соответственно видоизменяются и показатели качества классификации. В общем случае статистический критерий классификации может быть представлен в форме $\gamma(X) \lessgtr c$, где γ — известная функция X, а c — порог критерия. При изложении материала этого параграфа наряду с нейтральной математической терминологией будет использоваться терминология, «окрашенная» спецификой конкретных приложений.

1.2.1. Случай простого правила. Будем для удобства называть объекты первой совокупности «случаями» (случай брака, случай заболевания и т. п.), а объекты второй совокупности—«не-случаями». Пусть далее принимается гипотеза, что объект с характеристикой X является случаем, если $\gamma(X) < c$, и гипотеза, что объект является не-случаем, если $\gamma(X) \geqslant c$.

Результаты классификации изучаемой группы объектов удобно представить в виде табл. 1.1, в которой указано число объектов, удовлетворяющих условиям, наложенным на соответствующие строки и столбцы.

Таблица 1.1

Результат применения критерия	Статус объекта		Всего
	«случай» (H_1)	«не-случай» (H_2)	
Принимается гипотеза «случай» $\gamma(X) < c$	a	b	$a+b$
Отвергается гипотеза «случай» $\gamma(X) \geqslant c$	e	f	$e+f$
Всего	$a+e$	$b+f$	n

В практической (особенно медицинской) работе широко используют следующие характеристики, получаемые с помощью чисел, определенных в табл. 1.1.

Частота случаев — $(a + e)/n$.

Чувствительность критерия в обнаружении (предсказании) случая $a/(a + e)$, т. е. доля случаев, для которых $\gamma(X) < c$. С чувствительностью связано введенное ранее понятие ошибки первого рода (α) в проверке гипотезы, что изучаемый объект есть случай. Чувствительность $= 1 - \alpha$.

Специфичность критерия — $f/(b + f)$, т. е. доля не-случаев, для которых $\gamma(X) \geqslant c$. Специфичность равна $1 - \beta$, где β — ошибка второго рода в проверке гипотезы, что изучаемый объект случай.

Относительный риск — отношение вероятности быть случаем при условии, что гипотеза «случай» принята, к вероятности быть случаем при условии, что эта гипотеза отвергнута $R = \frac{a}{a+b} : \frac{e}{e+f}$.

Доля ложноположительных — $b/(a + b)$, т. е. доля не-случаев среди объектов, признанных случаями.

Доля ложноотрицательных — $e/(e + f)$, т. е. доля случаев среди объектов, признанных не-случаями.

Среди введенных характеристик только три независимых, остальные могут быть получены из них простым пересчетом. Представляется целесообразным выбрать в качестве ведущих частоту случаев (как параметр, связанный с выборочной схемой) чувствительность и специфичность (как параметры, связанные с разделимостью распределений случаев и не-случаев) или, что то же самое, частоту случаев и ошибки первого и второго рода. Никакие две из указанных характеристик не дают полного представления о ситуации. В прикладных исследованиях об этом часто забывают и сообщают только общий процент ошибочных диагностических заключений. При этом близость к нулю этого процента при низкой частоте случаев вообще не гарантирует высокую чувствительность критерия. Неполные наборы характеристик встречаются даже в высшей степени интересных работах [49, с. 262].

1.2.2. Изменение порога критерия. Часто в приложениях возникает необходимость описать качество классификации, достигаемое с помощью заданной функции $\gamma(X)$ при различных значениях c. Для этой цели достаточно привести одно число — частоту случаев и одну кривую — график «чувствительность — специфичность».

Предположим теперь дополнительно, что имеет место классическая модель Фишера (см. п. 1.1.2), в качестве $\gamma(X)$

используется логарифм отношения правдоподобия. В этом случае $\gamma(X)$ является линейной функцией X и, следовательно, для случаев и не-случаев $\gamma(X)$ имеет нормальное распределение с одной и той же дисперсией. Обозначим ее σ^2 и пусть

$$d = (\mathrm{E}(\gamma(X) \mid H_1) - \mathrm{E}(\gamma(X) \mid H_2))/\sigma. \qquad (1.37)$$

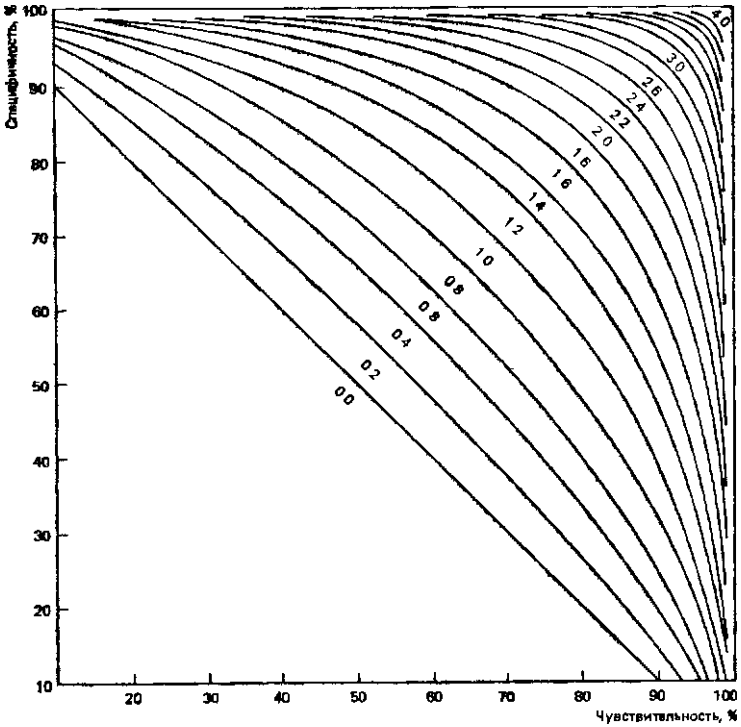


Рис. 1.3. Кривые «чувствительность — специфичность» для различных значений d (модель Фишера)

Тогда кривая «чувствительность — специфичность» (рис. 1.3) в параметрической форме имеет вид

$$(x(t),\ y(t)) \equiv (1 - \Phi(t-d),\ \Phi(t)), \qquad (1.38)$$

где $\Phi(t)$ — функция распределения стандартизованной нормальной величины. Если изменить масштабы по оси абсцисс и ординат по формулам $u = \psi(x)$, $v = \psi(y)$, где

ψ — функция, обратная к Φ, то кривая (1.38) перейдет в прямую

$$(d-t,\ t),\ -\infty < t < \infty. \qquad (1.38')$$

В самом деле,

$$u = \psi(1-\Phi(t-d)) = -\psi(\Phi(t-d)) = -(t-d);$$
$$v = \psi(\Phi(t)) = t$$

Существует специальная бумага, называемая *двойной нормальной*, на которой описанное выше преобразование выполнено. Кривые на ней распрямляются (рис. 1.4). Когда распределения $\gamma(X)$ для случаев и не-случаев по-прежнему нормальны, но имеют разные стандартные отклонения, кривая «чувствительность — специфичность» на двойной нормальной бумаге будет опять прямой, причем если φ — угол ее наклона к оси абсцисс, то отношение стандартного отклонения случаев к стандартному отклонению не-случаев равно $|\operatorname{tg}\varphi|$

Опыт показывает, что кривые «чувствительность — специфичность», построенные по реальным данным, при нанесении их на двойную нормальную бумагу часто распрямляются хотя бы в своей центральной части. Это дает возможность в интересующем исследователя диапазоне чувствительности (специфичности) характеризовать приближенно разделяющую силу используемого критерия $\gamma(X)$ одним числом d.

1.2.3. Условная вероятность быть случаем. В исследованиях, направленных на выявление риск-факторов стать за фиксированное время случаем, принято разбивать исходные объекты на несколько частей равного объема согласно увеличивающемуся риску стать случаем и для каждой части указывать соответствующую долю случаев [277, 322]. Если дополнительно предположить, что распределения для случаев и не-случаев приближенно нормальны с общей дисперсией, то по заданному значению d и частоте случаев легко найти распределение доли случаев для разбиения изучаемой популяции согласно риску быть случаем. В табл. 1.2 частота случаев указана для квартилей риска. Подобные таблицы можно использовать и в обратном направлении: по данной частоте случаев и долям случаев в квартилях (или децилях) найти соответствующее d. Аналогично, если при классификации используется трехградационное правило («объект является случаем», «отказ от классификации», «объект является не-случаем»), известны частоты принятия каждого из решений и соответствующие частоты ошибочных заключе-

Таблица 1.2

Частота	Квартиль	Расстояние d								
		0,2	0,4	0,6	0,8	1,0	1,2	1,4	1,6	1,8
0,01	1	0,0077	0,0057	0,0041	0,0028	0,0019	0,0012	0,0008	0,0005	0,0003
	2	0,0092	0,0082	0,0070	0,0057	0,0045	0,0034	0,0025	0,0018	0,0012
	3	0,0105	0,0106	0,0103	0,0096	0,0086	0,0075	0,0063	0,0050	0,0039
	4	0,0127	0,0156	0,0187	0,0218	0,0249	0,0278	0,0304	0,0327	0,0346
0,05	1	0,0387	0,0290	0,0211	0,0148	0,0101	0,0066	0,0042	0,0025	0,0015
	2	0,0462	0,0413	0,0357	0,0297	0,0238	0,0184	0,0137	0,0098	0,0067
	3	0,0523	0,0530	0,0519	0,0492	0,0452	0,0400	0,0342	0,0282	0,0224
	4	0,0628	0,0767	0,0913	0,1062	0,1209	0,1350	0,1497	0,1595	0,1694
0,10	1	0,0784	0,0597	0,0440	0,0314	0,0216	0,0143	0,0092	0,0056	0,0033
	2	0,0930	0,0840	0,0735	0,0622	0,0509	0,0401	0,0304	0,0221	0,0155
	3	0,1046	0,1064	0,1053	0,1015	0,0951	0,0865	0,0763	0,0651	0,0537
	4	0,1240	0,1500	0,1772	0,2049	0,2324	0,2591	0,2841	0,3071	0,3275
0,20	1	0,1611	0,1262	0,0959	0,0705	0,0501	0,0342	0,0225	0,0142	0,0087
	2	0,1880	0,1731	0,1556	0,1362	0,1158	0,0953	0,0757	0,0578	0,0424
	3	0,2087	0,2139	0,2156	0,2136	0,2079	0,1988	0,1864	0,1711	0,1535
	4	0,2422	0,2868	0,3329	0,3797	0,4262	0,4716	0,5154	0,5568	0,5954
0,50	1	0,4366	0,3745	0,3147	0,2585	0,2069	0,1607	0,1207	0,0873	0,0606
	2	0,4837	0,4670	0,4494	0,4306	0,4102	0,3878	0,3632	0,3364	0,3075
	3	0,5163	0,5330	0,5506	0,5694	0,5898	0,6122	0,6368	0,6636	0,6925
	4	0,5634	0,6255	0,6853	0,7415	0,7931	0,8393	0,8793	0,9127	0,9394

ний, то опять, зная общую частоту случаев в тех же предположениях о распределениях $\gamma(X)$ для случаев и не-случаев, можно оценить d. Верны и обратные утверждения для известных d и частоты случаев: 1) для заданных частот каждого из трех решений можно рассчитать соответствующие

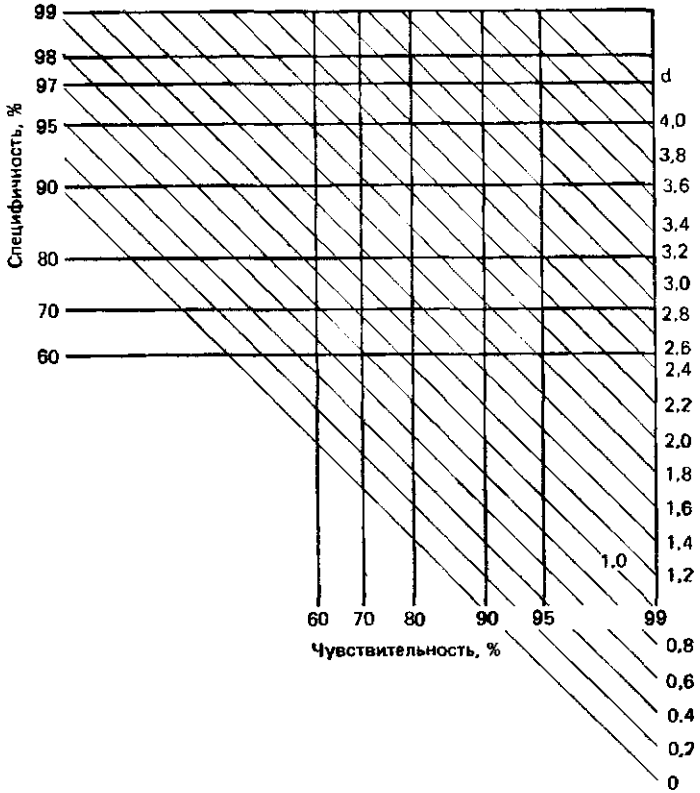


Рис. 1.4. Прямые «чувствительность — специфичность» для различных значений d (модель Фишера) на двойной нормальной бумаге

вероятности ошибок или 2) для заданных вероятностей ошибок найти частоты соответствующих решений.

Таким образом, при известной доле случаев и дополнительных предположениях о распределении $\gamma(X)$ величина d является очень удобной характеристикой разделения, позволяющей придать результатам классификации самую разнообразную форму: от графиков «чувствительность — специфичность» до доли случаев в квартилях риска и доли оши-

бочных заключений при заданном числе отказов от решения. Желательна определенная стандартизация представления результатов классификации. Величина d, определенная графически с помощью двойной нормальной бумаги, может служить универсальным эмпирическим параметром, характеризующим разделимость распределений.

1.2.4. Аналитические меры разделимости распределений. Пусть в модели Фишера (см. п. 1.1.2) d определено как в (1.37), тогда с учетом (1.16) и (1.17) из (1.12) получаем

$$d^2 = (M_2 - M_1)' \Sigma^{-1} (M_2 - M_1). \tag{1.39}$$

Для невырожденных многомерных нормальных распределений с общим Σ величина d^2, определенная формулой (1.39), называется *расстоянием Махаланобиса* между распределениями [16]. Она обладает следующими важными свойствами:

в задаче Фишера d однозначно определяет кривую «чувствительность — специфичность». При этом минимаксная ошибка классификации с помощью критерия отношения правдоподобия выражается формулой

$$\min_{\alpha} \max(\alpha, \beta) = \Phi(-d/2),$$

т. е. чем d больше, тем минимаксная ошибка меньше. При $d = 0$ ошибка равна 0,5 и соответствующие распределения совпадают;

если в задаче Фишера случайные векторы $Z_i' = (X_i', Y_i')$ $(i = 1, 2)$ состоят из двух взаимно независимых векторов X_i, Y_i, то

$$d^2(Z_1, Z_2) \quad d^2(X_1, X_2) + d^2(Y_1, Y_2). \tag{1.40}$$

Свойство (1.40) называют *аддитивностью* по отношению к независимым компонентам;

если $X_i \in N(M_i, \Sigma)$ $(i = 1, 2, 3)$ $|\Sigma| \neq 0$, то

$$d(X_1, X_3) \leqslant d(X_1, X_2) + d(X_2, X_3) \tag{1.41}$$

(*неравенство треугольника*).

В качестве обобщения расстояния Махаланобиса на произвольные распределения в теоретических работах широко используется *дивергенция* (в [91] *расхождение*) или, как еще иногда говорят, *расстояние Кульбака* между распределениями с плотностями f_i $(i = 1, 2)$

$$J = \int (f_1(X) - f_2(X)) \ln (f_1(X)/f_2(X))\, dX. \tag{1.42}$$

В модели Фишера $J = d^2$. Аналогично расстоянию Махаланобиса: $J = 0$ только тогда, когда распределения совпа-

дают; J также аддитивно по отношению к независимым компонентам и инвариантно относительно любого взаимно однозначного отображения координат. Какого-либо простого аналога (1.40) в литературе не приводится.

Другой мерой разделимости распределений является *расстояние Бхатачария* [160, гл. 9]

$$B = -\ln \int (f_1(X) \cdot f_2(X))^{1/2} \, dX. \tag{1.43}$$

Оно также инвариантно по отношению к взаимно однозначным отображениям координат, аддитивно по отношению к независимым компонентам, обращается в ноль при $f_1 = f_2$. В случае модели Фишера

$$B = d^2/8, \tag{1.44}$$

и в общем случае двух нормальных распределений (1.13)

$$B = \frac{1}{8}(M_2 - M_1)' \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (M_2 - M_1) + \\ + \frac{1}{2} \ln \frac{\left|\frac{1}{2}(\Sigma_1 + \Sigma_2)\right|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}}. \tag{1.45}$$

С помощью расстояния Бхатачария удается оценить сверху среднюю ошибку классификации при использовании критерия отношения правдоподобия

$$\delta \leqslant (\pi_1 \pi_2)^{1/2} \exp\{-B\}.$$

Более подробно эти вопросы обсуждаются в [189] и [160, гл. 3 и 9].

1.3. Два класса, заданные генеральными совокупностями

В параграфе рассматривается случай, когда классы описаны путем указания всех входящих в них наблюдений. Введенные ранее понятия — такие, как отношение правдоподобия, различные характеристики качества классификации, могут быть определены и в этом случае. Для описываемых ниже правил классификации подход с явным указанием возможных наблюдений более естествен. Эти правила (см. п. 1.3.2 — 1.3.5) не опираются прямо на отношение правдоподобия, но в простых случаях дают хорошее приближение к решающему правилу, построенному на его основе. Интерпретируемость полученных формул классификации часто служит

залогом их успешного применения. В этом смысле выделяются древообразные алгоритмы (см. п. 1.3.2), методы поиска характерных закономерностей (см. п. 1.3.4) и определение областей компетентности нескольких правил (см. п. 1.3.5).

1.3.1. Вычисление основных показателей. С целью большей преемственности обозначений с последующими главами зададим классы с помощью последовательности пар

$$(X_k,\ y_k),\ k=1,\ 2,\ \ldots,\ n\},\qquad (1.46)$$

где X_k — вектор возможного значения наблюдения, а $y_k = 1$, если X_k принадлежит первому классу, и $y_k = 2$, если второму. Будем считать также, что новые наблюдения извлекаются наудачу и независимо друг от друга из ряда (1.46). Таким образом, априорная вероятность гипотезы H_j $(j=1, 2)$ $\pi_j = P\{y=j\} = \sum\limits_{k:\,y_k=j} 1/n$; функция распределения X при гипотезе H_j будет $F(Z \mid H_j) = \sum\limits_{k:\,y_k=j,\ X_k<Z} 1/n_j$, где $n_j = \sum\limits_{k:\,y_k=j} 1$. Использованная при суммировании запись $X_k < Z$ означает, что для всех $1 \leqslant i \leqslant p$ имеем $x_k^{(i)} < z^{(i)}$. Отношение правдоподобия в точке X

$$\gamma(X) = \left(\sum_{k:\,y_k=2,\ X_k=X} 1/n_2\right) \Big/ \left(\sum_{k:\,y_k=1,\ X_k=X} 1/n_1\right). \qquad (1.1')$$

Ряд излагаемых ниже методов опирается на функции потерь, введенные в п. 1.1.4. Они легко могут быть оценены по ряду (1.46). Так, например, определенная с помощью формул (1.31) и (1.31′) вероятность ошибочной классификации запишется

$$\delta = \sum_{k=1}^{n} (y_k - \widehat{y}\,(X_k))^2/n. \qquad (1.31'')$$

В дальнейшем будем широко оперировать понятием условного математического ожидания, каждый раз подразумевая, что читатель самостоятельно может перейти от формулы типа (1.31′) к (1.31″).

1.3.2. Древообразные классификаторы. В основе описываемой ниже группы методов лежит понятие бинарного дерева — графа. Эти деревья принято изображать в перевернутом виде: корень — сверху, листья — внизу (рис. 1.5), при этом под словом «корень» понимаем самый верхний узел (вершину) графа, а под словом «лист» — узел, из которого

вниз не выходят стрелки к расположенным ниже узлам. С каждым узлом t дерева связаны следующие объекты:

R_t — подмножество пространства наблюдения (R);

V_t — подмножество генеральной совокупности (1.46) с $X_k \in R_t$;

A_t — правило классификации из разрешенного набора правил для $X \in R_t$.

Кроме того, для узлов «не-листьев» вводится правило разбиения R_t на два подмножества $R_{l(t)}$ и $R_{r(t)}$, таких, что $R_{l(t)} \cup R_{r(t)} = R_t$ и $R_{l(t)} \cap R_{r(t)} = \varnothing$.

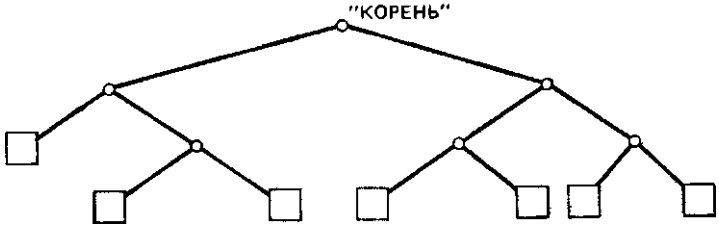


Рис. 1.5. Пример графа древообразного классификатора:
○ — промежуточный узел;
□ — концевой узел — «лист»

Древообразные классификаторы определяются рекурсивно. Для этого задаются: 1) критерий качества классификации на R_t; 2) разрешенный класс правил для построения A_t; 3) способ построения $R_{l(t)}$ и $R_{r(t)}$; 4) правило объявления узла листом, т. е. правило прекращения последующих делений.

В качестве корневого (нулевого) узла принимается узел с R_0, совпадающим со всем пространством возможных значений X, и $V_0 = \{(X_k, y_k), k = 1, \ldots, n\}$.

Критерий качества классификации. Пусть c_i — штраф за ошибочную классификацию, когда верна гипотеза H_i и функция потерь Q классификатора определена согласно (1.32′). Возьмем в качестве меры качества классификации на R_t классификатора A_t

$$Q(R_t, A_t) = \mathbf{E}(q^{A_t}(X) \mid X \in R_t). \qquad (1.47)$$

При этом чем Q меньше, тем классификатор лучше.

Разрешенный класс правил. Используются максимально простые классификаторы: линейные, линейные с дополнительными предположениями, например о структуре зависимостей признаков; простейшие, относящие все наблюдения

$X \in R_t$ к одной из классифицируемых совокупностей. Среди всех допустимых правил в качестве A_t выбирается правило, минимизирующее выбранную функцию риска. При этом ограничимся классами правил, на которых минимум достигается. Простота разрешенного класса правил компенсируется структурой дерева.

Правило разделения. Рассмотрим шкалу, в которой измерена j-я координата X ($j = 1, \ldots, p$). Если шкала номинальная, то элементарными событиями T будем называть события вида $x^{(j)} = a$ или $x^{(j)} \neq a$, где a — одно из возможных значений $x^{(j)}$. Если же шкала порядковая или интервальная, то элементарными событиями будем называть события вида $x^{(j)} \lesseqgtr a$. Рассмотрим теперь всевозможные разбиения R_t с помощью элементарных событий ($R_t \cap T$, $R_t \cap \overline{T}$), таких, что

$$\min\left(P\{X \in R_t \cap T\},\ P\{X \in R_t \cap \overline{T}\}\right) \geqslant \varepsilon_1, \qquad (1.48)$$

где ε_1 — некоторое заданное число. Тогда по определению критерия качества классификации (1.47)

$$Q(R_t,\ A_t) = P\{X \in T \mid X \in R_t\}\,\mathbf{E}\left(q^{A_t}(X) \mid X \in R_t \cap T\right) +$$
$$+ P\{X \in \overline{T} \mid X \in R_t\}\,\mathbf{E}\left(q^{A_t}(X) \mid X \in R_t \cap \overline{T}\right) \geqslant$$
$$\geqslant P\{X \in T \mid X \in R_t\} \inf \mathbf{E}\left(q^{A}(X) \mid X \in R_t \cap T\right) +$$
$$+ P\{X \in \overline{T} \mid X \in R_t\} \inf \mathbf{E}\left(q^{A}(X) \mid X \in R_t \cap \overline{T}\right). \qquad (1.49)$$

Нижняя грань в правой части (1.49) берется по всем допустимым классификаторам. Пусть T_t — одно из элементарных событий, удовлетворяющих (1.48), на котором достигается максимальная разность между левой и правой частями (1.49); положим тогда $R_{l(t)} = R_t \cap T_t$ и $R_{r(t)} = R_t \cap \overline{T}_t$.

Правило объявления узла листом. Узел t объявляется листом, если не существует элементарного события, удовлетворяющего условию (1.48), или такое событие существует, но

$$Q(R_t,\ A_t) - P\{X \in T_t \mid X \in R_t\}\,Q\left(R_{l(t)},\ A_{l(t)}\right) -$$
$$- P\{X \in \overline{T}_t \mid X \in R(t)\}\,Q\left(R_{r(t)},\ A_{r(t)}\right) \leqslant \varepsilon_2. \qquad (1.50)$$

Наглядный смысл условий (1.48), (1.50) очевиден: от усложнения классификатора должен быть заметный выигрыш и не следует слишком мельчить разбиения.

Древообразные классификаторы обладают рядом привлекательных свойств. Они относительно просты: при уменьшении ε_1 и ε_2 и соответствующем росте числа ветвей сходятся к классификатору, минимизирующему выбранную функцию потерь; инвариантны относительно монотонных преоб-

разований координат; легко интерпретируемы; при выборе в качестве допустимого класса простейших правил допускают наличие в классифицируемом векторе X ненаблюдаемых, так называемых «стертых», координат, если, конечно, эти координаты не используются в качестве аргументов ветвления.

В отечественной литературе древообразные правила называют также логическим методом классификации [94].

1.3.3. Метод потенциальных функций. Это — исторически один из первых достаточно универсальных методов построения классификационных правил в условиях хорошей разделимости классов [13, 131]. Он представляет собою пример результата научного направления, центр тяжести которого лежит не в оптимальном решении задачи классификации при дефиците выборочной информации, а в разработке рекуррентной процедуры, удобной для ЭВМ и дающей решение в условиях большой выборки. Метод основан на предположении, что объекты с близкими значениями X принадлежат одному классу. Поэтому при классификации нового объекта X надо лишь подсчитать «относительные потенциалы» в X, порожденные объектами первого и второго классов, и отнести объект к тому классу, чей относительный потенциал выше. Более точно: пусть в пространстве наблюдений определено расстояние ρ, например евклидово. Относительный потенциал в X, созданный объектами j-го класса, подсчитывается как

$$\Phi_j(X) = \sum_{k:y_k=j} \varphi(\rho(X, X_k))/n_j, \tag{1.51}$$

где $\varphi(u)$ — некоторая известная положительная функция положительного аргумента, стремящаяся к нулю при $u \to \infty$, например $\exp\{-\alpha\rho^\beta\}$ или $(1+\alpha\rho^\beta)^{-1}$, где $\alpha > 0$, $\beta > 0$. Правило классификации записывается:

$$\Phi_2(X) \gtrless \Phi_1(X) \Rightarrow \begin{cases} H_2 \\ H_1 \end{cases}. \tag{1.52}$$

Изложенный алгоритм близок с описываемым в § 3.3 непараметрическим методом классификации. Различие заключается в том, что статистическому подходу более соответствовало бы использование: 1) вместо относительного потенциала в X абсолютного, как лучше учитывающего априорные вероятности классов и 2) вместо разности (1.52) отношения

$$\frac{\Phi_2(X)}{\Phi_1(X)} \gtrless c \Rightarrow \begin{cases} H_2 \\ H_1 \end{cases}. \tag{1.53}$$

1.3.4. Поиск характерных закономерностей. Ниже описывается общая логическая схема одного из наиболее известных алгоритмов, возникшего из эвристических соображений о деятельности человека при распознавании образов — алгоритма «Кора» [28, 43]. В нем рассматриваются все возможные конъюнкции вида

$$T_{i_1} \cap T_{i_2} \cap \ldots \cap T_{i_l} (l \leqslant l_0), \qquad (1.54)$$

где T — события, определенные в п.1.3.2 при введении правил разделения, а l_0 — некоторое наперед заданное число (в алгоритме «Кора» $l_0 = 3$). Среди конъюнкций выделяются те, которые характерны (верны на обучающей выборке чаще, чем некоторый порог $1 - \varepsilon_1$) для одного из классов и не характерны для другого (верны реже, чем в доле случаев ε_2 (в алгоритме «Кора» $\varepsilon_2 = 0$)). Если коэффициент корреляции между какими-либо двумя выделенными конъюнкциями по модулю более $1 - \varepsilon_3$, то оставляется «наилучшая» с точки зрения различения классов из них, а если конъюнкции эквивалентны, то более короткая (имеющая в представлении (1.54) меньшее l) или просто отобранная ранее. Параметры ε_1, ε_2, ε_3 подбираются так, чтобы общее число отобранных (информативных) конъюнкций не превосходило некоторого числа m. Для нового наблюдения X подсчитывается m_i — число характерных для i-го класса отобранных конъюнкций, которые верны в точке X. Если $m_1 > m_2$, то принимается решение, что верна гипотеза H_1, в противном случае — что верна гипотеза H_2. Поскольку при отборе конъюнкций в принципе возможен полный перебор, вычислительный процесс должен быть организован эффективно, чтобы не рассматривать бесперспективные ветви. Алгоритм «Кора» зарекомендовал себя удачным в ряде прикладных областей [28, 82]. Идея поиска закономерностей, характерных для одного из классов, положена в основу алгоритма автоматизированного поиска гипотез [49].

1.3.5. Коллективы решающих правил. В прикладных исследованиях для классификации наблюдений иногда одновременно используется не одно, а несколько решающих правил. При этом, естественно, встает вопрос о выборе правила объединения частных заключений и как ответ на него возникает двухуровневая схема принятия решения (рис. 1.6). На рисунке первый уровень — блоки $A_1 - A_L$, второй — блок синтеза. В изложенном выше алгоритме «Кора», если считать отдельными правилами отобранные конъюнкции, решение принимается большинством «голосов», осуществившихся при X конъюнкциях. В системах, в которых используются высокоспецифические (см. п. 1.2.1) правила риска, объеди-

нение возможно по принципу «максимум предсказанного риска». В [131] предложен третий подход — выделение областей компетентности каждого из использованных алгоритмов. Пусть функция потерь $Q(R, A)$ правила A на множестве R определена по формуле (1.47). Правило A_{l^*} $(1 \leqslant l^* \leqslant L)$ считается на R компетентным среди правил A_l $(1 \leqslant l \leqslant L)$, если

$$Q(R, A_{l^*}) = \min_l Q(R, A_l).$$

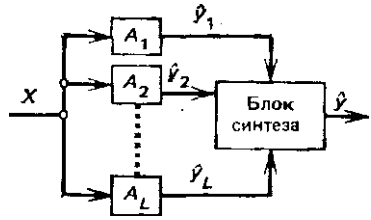


Рис. 1.6. Двухуровневая схема принятия решения

При классификации сначала выбирается наиболее компетентный алгоритм, затем с его помощью принимается решение (рис. 1.7, где F — блок выбора наиболее компетентного алгоритма).

Пусть пространство наблюдений R разбито на L областей R_l, таких, что $R = \bigcup R_l$ и $R_l \cap R_{l'} = \varnothing$ при $l \neq l'$ и для каждого R_l указан компетентный на нем алгоритм A_{l^*}. Тогда на первом шаге по значению X находится $R_l : X \in R_l$

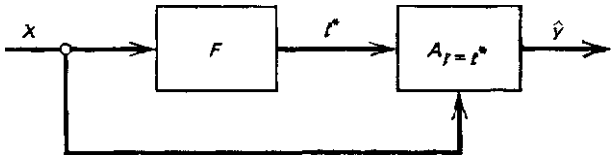


Рис. 1.7. Схема классификации с использованием областей компетентности

и по R_l — номер компетентного на нем алгоритма l^*. На втором шаге правило A_{l^*} применяется к X. В качестве примера изложенного подхода можно указать древообразные классификаторы (см. п. 1.3.3), где механизм нахождения R_l - - ветви дерева (без листьев), правило решения A_{l^*} — правило соответствующего листа.

Опишем два простейших правила нахождения областей компетентности. *Метод априорного разбиения* пространства наблюдения R на подобласти основан на профессиональных соображениях конкретной науки. Для каждой из введенных подобластей находят компетентное на ней правило классификации. Этот метод удобен тем, что введенные области легко строить и интерпретировать. Во втором методе области ком-

петентности $\{R_l\}$ строятся *локально* путем построения алгоритма, с помощью которого для каждого наблюдения можно вычислить, какой области R_l оно принадлежит. Пусть для каждого X_0 можно ввести семейство вложенных друг в друга расширяющихся окрестностей. Фиксируем какое-либо число k и для X_0 найдем наименьшую окрестность $O(X_0)$, которая включает в себя не менее k точек последовательности (1.46). Пусть далее $c_{li} = 1$, если i-е наблюдение последовательности классифицировано A_l алгоритмом правильно, и $c_{li} = 0$ в противном случае; $c_l = \Sigma c_{lm}$, где суммирование проводится по всем точкам, принадлежащим $O(X_0)$, и

$$c_{l^*} = \max_l c_l. \tag{1.55}$$

Тогда точка X_0 объявляется принадлежащей области компетентности правила A_{l^*}. Чтобы обойти случаи, когда максимум в (1.55) достигается не на одном, а на нескольких значениях $l_{i_1}, \ldots, l_{i_m}$, положим l^* равным наименьшему из них. Если на R определено расстояние между точками ρ, то окрестности можно задавать с помощью расстояний и в определении c_{li} вместо 1 брать $g(\rho(X_i, X_0))$, где g — некоторая убывающая функция от положительного аргумента. Например, $g(\rho) = 1/(a + \rho)$, где $a > 0$. В [131] предлагается для выделения областей компетентности использовать также метод потенциальных функций.

1.4. Отбор информативных переменных

Любое практическое исследование с применением методов статистической классификации включает в себя в виде специального этапа отбор информативных для классификации переменных. Дело здесь заключается не столько в экономии затрат на сбор не- или малоинформативных признаков, сколько в том, как увидим в следующей главе, что включение в решающее правило в условиях дефицита выборочной информации малоинформативных признаков ухудшает (!) среднюю эффективность классификации. В этом параграфе рассматриваются два принципиально отличных подхода к отбору переменных. В первом из них делаются сильные математические предположения о характере классифицируемых распределений и это позволяет четко и однозначно ответить на вопросы, следует или нет включать рассматриваемую переменную в решающее правило и если нет, то почему. Во втором подходе специальных предположений не делается, предла-

гаются некоторые эвристические итеративные процедуры, каждый шаг которых разумен, но общий результат их применения осмыслить и изучить трудно.

1.4.1. Модель Фишера с дополнительными предположениями о структуре зависимостей признаков. Рассмотрим сначала простейшую математическую модель двух нормальных распределений с независимыми переменными

$$N_1 = N\left(\begin{pmatrix} m_1^{(1)} \\ \dots \\ m_1^{(p)} \end{pmatrix}, \begin{bmatrix} \sigma_1^2 \dots 0 \\ \dots\dots \\ 0 \dots \sigma_p^2 \end{bmatrix}\right) \text{ и}$$

$$N_2 = N\left(\begin{pmatrix} m_2^{(1)} \\ \dots \\ m_2^{(p)} \end{pmatrix}, \begin{bmatrix} \sigma_1^2 \dots 0 \\ \dots\dots \\ 0 \dots \sigma_p^2 \end{bmatrix}\right).$$

Решающее правило и расстояние Махаланобиса между N_1 и N_2 согласно (1.12), (1.39) имеют вид

$$h(X) = \Sigma\,(x^{(i)} - (m_1^{(i)} + m_2^{(i)})/2)\,(m_2^{(i)} - m_1^{(i)})\,\sigma_i^{-2} \gtrless c; \qquad (1.56)$$

$$d^2(N_1, N_2) = \Sigma\,(m_2^{(i)} - m_1^{(i)})^2\,\sigma_i^{-2}. \qquad (1.57)$$

Естественно считать неинформативными переменные, у которых не отличаются средние, т. е. соответствующие $m_2^{(i)} - m_1^{(i)} = 0$, и малоинформативными переменные, у которых $(m_2^{(i)} - m_1^{(i)})^2\sigma_i^{-2} \leqslant \varepsilon$, где ε — некоторое число. Таким образом, в простейшей математической модели об информативности переменной можно судить по ее одномерным распределениям при H_1 и H_2. В общем случае это неверно, так как даже переменные, имеющие идентичные одномерные распределения при H_1 и H_2, могут нести существенную информацию о проверяемых гипотезах в силу взаимозависимости переменных. В качестве примера вернемся к рис. 1.1. Распределения $x^{(2)}$ при обеих гипотезах совпадают, однако эта переменная в совокупности с $x^{(1)}$ существенна для классификации.

Рассмотрим теперь модель Фишера с древообразной структурой зависимостей (ДСЗ) переменных [12, п. 4.2.3] $N_1 = N(M_1, \Sigma)$ и $N_2 = N(M_2, \Sigma)$, где Σ имеет ДСЗ. Внедиагональные элементы $\Sigma^{-1} = \|\sigma^{ij}\|$ отличны от нуля тогда, когда они принадлежат G — графу структуры зависимостей распределений. На основании (1.12)

$$h(X) = \sum_i (x^{(i)} - (m_1^{(i)} + m_2^{(i)})/2)\,\sigma^{ii}\,(m_2^{(i)} - m_1^{(i)}) +$$

$$+ \sum_{(i,\,j)\in G} (x^{(i)} - (m_1^{(i)} + m_2^{(i)})/2)\,\sigma^{ij}\,(m_2^{(j)} - m_1^{(j)}).$$

В последнюю сумму наряду с парой (i, j) входит и (j, i). Таким образом, в $h(X)$ входят только те переменные, для которых или 1) их индивидуальный вклад в разделение отличен от нуля, т. е. $m_2^{(i)} - m_1^{(i)} \neq 0$, или 2) индивидуальный вклад равен нулю, но они непосредственно связаны на графе структуры зависимостей с переменными j, для которых $m_2^{(j)} - m_1^{(j)} \neq 0$.

Этот результат остается верным и для распределений с $R(k)$-зависимостью [12, 4.4].

1.4.2. Функции потерь. Пусть $S \subseteq \{1, \ldots, p\}$ — некоторое подмножество координат X. Обозначим $X^{(S)}$ входящие в него компоненты X. Для того чтобы произвести отбор информативного подмножества координат, вводится функция потерь $Q(S)$, обладающая следующими свойствами:

$$\text{для } S_1 \subseteq S \quad Q(S_1) \geqslant Q(S); \tag{1.58}$$

для $Q(S_1) = Q(S)$ наборы признаков S_1 и S считаются эквивалентными,

для $Q(S_1) > Q(S)$ набор признаков S считается предпочтительнее набора S_1.

В качестве $Q(S)$ можно взять, например, ожидаемую ошибку байесовского классификатора

$$Q_B(S) = \mathbf{E}\left(1 - \max_i \widehat{\pi}_i(X^{(S)})\right), \tag{1.59}$$

где $\widehat{\pi}_i(X^{(S)})$ — условная вероятность гипотезы H_i при наблюдении $X^{(S)}$, а математическое ожидание берется по мере $P(X) = \Sigma \pi_i P(X \mid H_i)$. Нетрудно показать, что для Q_B условие (1.58) выполняется. Это условие помогает эффективно организовать процесс отбора признаков. Пусть, например, исследованы два подмножества признаков S_1 и S_2 и пусть $Q(S_1) < Q(S_2)$, тогда нет необходимости исследовать любое из подмножеств S_2, так как в силу (1.58) заранее известно, что они менее предпочтительны, чем S_1. Это соображение легло в основу многих высокоэффективных вычислительных алгоритмов.

Заметим, что для определения Q можно использовать любую из мер разделимости распределений, введенных в п. 1.1.5. Для этого достаточно положить

$$Q(S) = -m(F_1(X^{(S)}),\ F_2(X^{(S)})),$$

где m — соответствующая мера, а $F_i(X^{(S)})$ — функция распределения $X^{(S)}$ при условии, что верна гипотеза H_i $(i = 1, 2)$. Взаимоотношения между различными функциями потерь систематизированы в [189]. Для нас только важно, что нет функции потерь, которая отбирала бы признаки

в том же порядке, что и Q_B. Вместе с тем многочисленные примеры показывают, что корреляция между наборами, отобранными с помощью различных функций Q, высокая.

В заключение отметим, что только в сильных предположениях п.4.5.1 удается надежно оценивать групповую информативность признаков по индивидуальной. В общем случае неожиданности возможны даже в модели Фишера. В [325] утверждается, что для любого набора чисел $\{Q(S): S \subset \{1, \ldots, p\}\}$, удовлетворяющего условиям согласования (1.58), можно подобрать задачу Фишера, в которой числу i соответствует вектор $X^{(i)}$ и при этом для всех S имеем $Q_B(S) = Q(S)$. В частности, возможна, например, такая неожиданная комбинация информативностей, когда при трех группах признаков индивидуально наиболее информативна третья группа, а попарно — совокупность первых двух групп.

1.4.3. Схемы последовательного испытания наборов признаков. Общая логическая схема рассуждения здесь традиционна:

выбирается функция потерь $Q(S)$;

для каждого набора переменных, порождаемого с помощью какой-либо пошаговой процедуры, строится наилучший (в смысле Q) критерий классификации;

среди всех построенных наборов отбирается тот (те), в который входит наименьшее число переменных и при котором Q минимально.

Схемы генерации наборов переменных, по существу, аналогичны схемам, используемым при отборе переменных в регрессионном анализе [12, п. 8.7.4] и опираются на эвристическое предположение, что наилучший набор из $k+1$ переменных часто содержит в себе наилучший набор из k переменных. Однако в общем случае так же, как и в регрессии, это предположение неверно, и пошаговые процедуры не гарантируют получения оптимального набора переменных, т. е. в общем случае без дополнительных предположений полный перебор неизбежен (см. п.1.4.2). Практические аспекты отбора переменных в условиях дефицита выборочной информации обсуждаются во второй и третьей главах.

1.5. Три и более полностью определенных класса

1.5.1. Общая постановка задачи. Пусть имеется k генеральных совокупностей (классов) с плотностями распределения вероятностей $f_j(X)$,

где $j = 1, \ldots, k$; H_j — гипотеза, состоящая в том, что наблюдение X извлечено из j-й совокупности; π_j — априорные вероятности гипотез; $c(j \mid i)$ — цена ошибочной классификации наблюдения из i-й совокупности как наблюдения из j-й совокупности. Задача в том, чтобы разбить пространство наблюдений R на k попарно непересекающихся областей $R_1, \ldots, R_k$, таких, что если $X \in R_j$, то принимается гипотеза H_j $(j = 1, \ldots, k)$, и при этом минимизировать потери. Вероятность принять гипотезу H_j, когда X извлечено из i-й совокупности, составит

$$P(H_j \mid H_i) = \int_{R_j} f_i(X)\, dX \tag{1.60}$$

и функция потерь

$$Q = \sum_{i=1}^{k} \pi_i \left[\sum_{j \neq i} c(j \mid i) P(H_j \mid H_i) \right]. \tag{1.61}$$

Если значение вектора X фиксировано, то $\widehat{\pi_i}(X)$ — апостериорная вероятность H_i — равна

$$\widehat{\pi_i}(X) = \frac{\pi_i f_i(X)}{\sum_j \pi_j f_j(X)}, \tag{1.62}$$

а ожидаемые потери при решении, что X извлечен из j-й совокупности, составят

$$\sum_{\substack{i=1 \\ i \neq j}}^{k} \frac{\pi_i f_i(X)}{\sum_l \pi_l f_l(X)} c(j \mid i). \tag{1.63}$$

Очевидно, что потери будут минимальными, если

$$\sum_{i \neq j} \pi_i f_i(X) c(j \mid i) = \min_l \sum_{i \neq l} \pi_i f_i(X) c(l \mid i). \tag{1.64}$$

Поэтому определим R_j как множество точек, для которых верно (1.64). Если минимум для некоторого X достигается при нескольких значениях j, то относим X к любому из соответствующих R_j. Сформулированное правило при

$$c(j \mid i) \equiv 1 \tag{1.65}$$

очевидно сводится к отнесению X к тому R_j, для которого $\pi_j f_j(X)$ наибольшее. Это правило классификации называют байесовским.

В случае, когда распределения генеральных совокупностей непрерывны с точностью до значений X, попадание в

которые имеет нулевую вероятность, R_j для байесовского классификатора могут быть также определены как

$$R_j = \{X : h_{ji} = \ln (f_j(X)/f_i(X)) > -\ln (\pi_j/\pi_i) \text{ для всех } i \neq j\}, \tag{1.66}$$

т. е. построение байесовского классификатора в случае k классов сводится к последовательному построению байесовских классификаторов для двух классов, т. е. к методам, которые уже рассмотрены в предшествующих параграфах.

1.5.2. Модель нескольких многомерных нормальных распределений с общей ковариационной матрицей. Эту модель иногда называют также моделью Фишера для k классов. Пусть $N(M_i, \Sigma)$ — распределение X в i-м классе ($i = 1, \ldots, k$). Тогда согласно (1.64) и (1.66) области R_i определяются условиями

$$R_j = \{X : h_{ji} = (X - (M_i + M_j)/2)' \Sigma^{-1} (M_j - M_i) \geqslant \geqslant \ln (\pi_i/\pi_j),\ i = 1, \ldots, k;\ i \neq j\}. \tag{1.67}$$

Поскольку каждая из функций h_{ji} линейна по X, то область R_j ограничена гиперплоскостями.

Покажем теперь, как оценить вероятности правильной классификации. Для этого рассмотрим случайную величину $h_{ij}(X)$. При гипотезе H_i она как линейная функция нормально распределенных величин имеет распределение $N\left(\frac{1}{2} d_{ij},\ d_{ij}^2\right)$, где

$$d_{ij} = d_{i,jj} = (M_i - M_j)' \Sigma^{-1} (M_i - M_j). \tag{1.68}$$

Ковариации между $h_{ij}(X)$ и $h_{ik}(X)$ при гипотезе H_i равны [16, § 6.7]

$$d_{i,jk} = (M_i - M_j)' \Sigma^{-1} (M_i - M_k). \tag{1.69}$$

Таким образом, распределение векторов $\Gamma_i(X) = (h_{i1}(X), \ldots, h_{i,i-1}(X), h_{i,i+1}(X), \ldots, h_{ik}(X))'$ является многомерным нормальным с вектором средних $M_{\Gamma_i} = \frac{1}{2}(d_{i1}, \ldots, d_{i,i-1}, d_{i,i+1}, \ldots, d_{ik})'$ и ковариационной матрицей $\Sigma_{\Gamma_i} = \|d_{i,jl}\|_{j,l \neq i}^{k}$. Для нахождения вероятности правильной классификации вектора $\Gamma_i(X)$ надо найти вероятность попадания его в область (1.67).

1.5.3. Упорядоченные классы. Иногда между введенными в п. 1.3.1 классами можно ввести отношение предшествования ($\prec$). Если это отношение транзитивно, т. е. если для любых классов i, j, k из $i \prec j$ и $j \prec k$ следует, что $i \prec k$,

то классы будем называть *упорядоченными*. Упорядочение может быть связано с содержательным истолкованием классов и с их геометрическим расположением вдоль какой-либо гладкой кривой в выборочном пространстве. В случае, когда классы соответствуют последовательным стадиям некоторого процесса, содержательное и геометрическое упорядочения часто совпадают.

При работе с упорядоченными классами используется следующий методический прием. С каждым классом i связывают волевым путем выбранное число θ_i так, чтобы разности $\theta_i - \theta_j$ соответствовали интуитивному представлению исследователя о «расстоянии» между классами i и j, и находят функцию от наблюдения $t(X)$, такую, чтобы разность $\theta - t(X)$ была бы в некотором смысле наименьшей. Классификацию далее осуществляют в зависимости от значения $t(X)$.

В одной из конкретных реализаций этого приема [24] на распределение $t(X)$ в классах накладывается ограничение

$$\mathbf{E}(t(X) \mid H_i) = \theta_i. \tag{1.70}$$

Качество классификации измеряется как

$$v(t) = \sum_{1}^{k} w_i \mathbf{D}(t(X) \mid H_i), \tag{1.71}$$

где $w_i > 0$, $\Sigma w_i = 1$. Функция t, минимизирующая (1.71) при условиях (1.70), имеет вид

$$t(X) = \frac{\mathbf{f}'(X)\,\mathbf{B}^{-1}\boldsymbol{\theta}}{\mathbf{w}'\,\mathbf{f}(X)}, \tag{1.72}$$

где $\mathbf{w} = (w_1, \ldots, w_k)'$; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$; $\mathbf{f}(X) = (f_1(X), \ldots, f_k(X))'$, где $f_i(X)$ — плотность распределения X при H_i; $\mathbf{B} = ||b_{ij}||$, где $b_{ij} = \int \frac{f_i(X) f_j(X)}{\mathbf{w}'\mathbf{f}(X)}\, dX$. Если рассматривать w_i как априорную вероятность того, что наблюдение выбрано из i-го класса (H_i), то $v(t)$ — среднее квадратическое отклонение $t(X)$ от соответствующего θ, а $t(X)$ — линейная функция от апостериорных вероятностей классов $w_i f_i(X)/\mathbf{w}'\mathbf{f}(X)$. Если классы не пересекаются, т. е. при $i \neq j$ $f_i(X) \cdot f_j(X) = 0$, то $v(t) = 0$ и функция на каждом из классов равна соответствующему значению θ.

В условиях дефицита выборочной информации о распределениях к предположениям типа (1.70), (1.71) иногда добавляют предположения, что $f_i(X)$ нормальны и их средние лежат на одной прямой.

ВЫВОДЫ

1. Среди критериев классификации в одно из двух известных распределений с заданной ошибкой первого рода α наименьшую ошибку второго рода β имеет критерий отношения правдоподобия вида (1.1). *Байесовский классификатор* определяется с помощью формулы (1.2). Он минимизирует вероятность ошибочной классификации. При выборе между двумя многомерными нормальными распределениями с общей ковариационной матрицей (*модель Фишера*) граница критической области критерия является гиперплоскостью в пространстве наблюдений, зависящей от параметров распределений по формуле (1.12). Наряду с критерием отношения правдоподобия на практике широко используются правила классификации, критические области которых находятся путем минимизации заданной *функции потерь* при данных ограничениях на границу критической области. При этом функцию потерь и ограничения на границу критической области обычно выбирают так, чтобы в случае, когда верна одна из базовых теоретических моделей классификации, построенный критерий совпадал с критерием отношения правдоподобия.

2. Для характеристики простого правила классификации при двух классах в условиях полностью известных распределений необходимо использовать не менее двух чисел — вероятностей ошибок α и β. К ним часто добавляют третье число — вероятность того, что наблюдение извлечено из одного из классов. Все остальные характеристики правила получаются простым пересчетом из указанных трех базовых.

На практике широко используется прием, когда классификация проводится с переменным порогом и для каждого диапазона значений отношения правдоподобия указывается условная вероятность, что наблюдение принадлежит одному из классов при условии, что оно попало в данный диапазон. В этом случае в качестве базовой характеристики критерия рассматривается кривая $(1 - \alpha(c),\ 1 - \beta(c))$, где c — порог критерия. Ее называют *кривой «чувствительность — специфичность»*. В модели Фишера при специальном выборе масштаба на координатных осях все кривые «чувствительность — специфичность» превращаются в параллельные прямые, идущие под углом 135° к оси абсцисс и отстоящие от прямой $(u,\ 1 - u)$ на расстояние, пропорциональное d, где d^2 — расстояние Махаланобиса, определенное формулой (1.39).

3. Наряду с аналитическим описанием распределений в классах используется также прием задания распределений путем указания соответствующих генеральных совокупностей. Его можно рассматривать как теоретическое представление большой выборки. Все основные показатели распределений могут быть оценены и в этом случае. Вместе с тем прямое задание генеральных совокупностей позволяет использовать при классификации методы, осуществление которых невозможно или крайне затруднительно при аналитическом задании распределений. Одним из примеров здесь являются *древообразные* или *логические классификаторы*.

Они обладают рядом привлекательных свойств: просты, легко интерпретируемы, при увеличении числа ветвей сводятся к классификатору, минимизирующему заданную функцию потерь.

4. При построении классификационного правила часто производится отбор информативных для разделения классов координат. При этом используются два методических подхода. В первом из них на взаимозависимость переменных накладываются сильные упрощающие предположения, но сам отбор не требует чрезмерных вычислений, и всегда можно ответить на вопрос, почему берется или отвергается переменная. Второй подход связан с минимизацией некоторой функции потерь и проводится путем последовательного испытания наборов признаков. При этом широко используются различные эвристические соображения, направленные на то, чтобы уменьшить перебор. Они часто хорошо оправдываются на практике, однако встречают серьезные теоретические возражения. Четкого ответа на вопрос, почему включена или отвергнута переменная, при втором подходе дать нельзя.

5. В случае $k > 2$ классов построение байесовского классификатора сводится к построению байесовских классификаторов для всех пар классов. Наиболее распространенная модель в этом случае — это предположение, что $F_i = N(M_i, \mathbf{\Sigma})$, где матрица $\mathbf{\Sigma}$ одна и та же для всех классов. Особый интерес представляет случай, когда можно предположить, что классы упорядочены по какому-либо признаку. В этом случае каждому классу приписывается некоторое число θ так, чтобы расстояние между последовательными числами отвечало интуитивной идее исследователя о расстоянии между классами.

Далее строится $t(X)$ оценка θ для наблюдения X, и классификация проводится по величине t.

Глава 2. ТЕОРЕТИЧЕСКИЕ РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ ПРИ НАЛИЧИИ ОБУЧАЮЩИХ ВЫБОРОК (ДИСКРИМИНАНТНЫЙ АНАЛИЗ)

В предыдущей главе распределения векторов X внутри классов предполагались известными: они задавались аналитически или с помощью перечисления всех возможных значений X. С использованием этой информации строилось правило (критерий) классификации. В этой и последующих двух главах распределения X внутри классов определяются лишь частично. При этом используются два вида информации: предположения о свойствах распределений (гладкость, принадлежность к некоторому известному параметрическому классу) и обучающая выборка. Совокупность алгоритмов, порождающих на основании предположений и выборки конкретное правило классификации, называют дискриминантным анализом (ДА). Построенное правило классификации как функция от случайной выборки отражает ее особенности и тоже в определенной степени случайно. Это затрудняет сравнение алгоритмов ДА.

Цель главы — познакомить с основными понятиями ДА, методами сравнения алгоритмов и результатами теоретического исследования свойств алгоритмов в условиях дефицита выборочной информации.

2.1. Базовые понятия дискриминантного анализа

2.1.1. Выборка, предположения, алгоритм, оценка качества дискриминации. В дальнейшем предполагается, что случайная *выборка* представляет собой последовательность независимых пар наблюдений вида (1.46)

$$W_n = \{(X_i, y_i),\ i = 1, \ldots, n\}, \tag{2.1}$$

где $P\{y_i = j\} = \pi_j$, $j = 1, \ldots, k$; y_i трактуется как номер класса, которому принадлежит наблюдение X_i; π_j — неизвестная вероятность, что X будет извлечено из j-го класса; число классов k известно исследователю, $\Sigma\pi_j = 1$; все X_i принадлежат одному и тому же пространству наблюдений; X_i — такие, что $y_i = j$, одинаково распределены с неизвестной исследователю функцией распределения F_j. Чи-

сло $y_i = j$ в выборке будем обозначать n_j и называть объемом выборки из j-го класса. *Предположения* о характере распределений F_j ($j = 1, \ldots, k$) в наиболее информативном случае утверждают, что F_j принадлежат некоторому известному семейству распределений, зависящему от неизвестного векторного параметра Θ.

В модели Фишера предполагается, что $X \in R^p$. Имеются всего два класса ($k = 2$), F_j ($j = 1, 2$) имеют многомерные нормальные распределения с общей невырожденной ковариационной матрицей. В этом случае каждое из F_j определяют одним p-мерным вектором средних, своим для каждого распределения, и $p(p+1)/2$ — параметрами ковариационной матрицы, общими для обоих распределений.

Иногда предполагается с точностью до неизвестных параметров аналитический вид отношения правдоподобия и, наконец, самый слабый вид предположений — постулирование непрерывности F_j. Формальная процедура, использующая часть информации предположений и выборку для получения конкретного классификационного правила, называется *алгоритмом*. Приведем примеры описания алгоритмов.

Пример 2.1. Имя: *Подстановочный алгоритм с независимой оценкой параметров* ($k = 2$). Применяется: в случаях, когда предполагается, что

1) $F_j(X) \quad F(X, \Theta_j)$, $j = 1, 2$;

2) Θ_1 и Θ_2 не имеют общих значений координат.

Вычисления над выборкой: независимо для каждого из Θ_j строятся оценки максимального правдоподобия $\widehat{\Theta}_j$ [12, гл. 8]. При этом при оценке Θ_1 используются n_1 наблюдений из первого класса, при оценке Θ_2 — n_2 наблюдений из второго.

Прогностическое правило:

$$\frac{f_2(x, \widehat{\Theta}_2)}{f_1(x, \widehat{\Theta}_1)} \gtrless c \Rightarrow \begin{cases} H_2 \\ H_1 \end{cases},$$

где f_j — плотности соответственно распределений F_j ($j = 1, 2$) и H_j — гипотеза о том, что новое наблюдение извлечено из j-го класса.

Пример 2.2. Имя: *Подстановочный алгоритм в задаче Фишера*. Применяется: в случаях, когда предполагается, что

$$F_j = N(M_j, \Sigma), \; j = 1, 2, \; |\Sigma| > 0,$$

M_j и Σ неизвестны. В отличие от предшествующего примера матрица Σ — общая для обоих распределений.

Вычисления над выборкой: строятся оценки максимального правдоподобия для M_1, M_2, Σ:

$$\overline{X}_j = \sum_{i\ y_i = j} X_i / n_j \ (j = 1, 2); \tag{2.2}$$

$$S = \sum_{j=1}^{2} \sum_{i:y_i=j} (X_i - \overline{X}_j)(X_i - \overline{X}_j)' / (n_1 + n_2 - 2). \tag{2.3}$$

Прогностическое правило:

$$\frac{f(X, \overline{X}_2, S)}{f(X, \overline{X}_1, S)} \gtrless c \Rightarrow \begin{matrix} H_2 \\ H_1 \end{matrix}, \tag{2.4}$$

где $f(X, \overline{X}, S) = (2\pi)^{-p/2} |S|^{-1/2} \exp\{-(X - \overline{X})' S^{-1} \times$

$\times (X - \overline{X})/2\}$.

Оценка качества построенного правила классификации – завершающая операция ДА. В ней используются оценки определенных в гл. 1 показателей качества разделения. Оценка качества дискриминации — это не только оценка конкретного правила классификации, но в более широком смысле и проверка удачности сделанных предположений и выбора алгоритма ДА.

В гл. 1 объектом изучения были различные правила классификации. В настоящей главе — алгоритмы, порождающие конкретные правила. В приведенных выше описаниях $\widehat{\Theta}_j$ $(j = 1, 2)$ случайны, следовательно, случайны и правила.

2.1.2. Основные виды ошибок. Базовым понятием, как и в гл. 1, остается вероятность ошибочной классификации конкретного правила. Теперь, однако, это уже случайная величина, зависящая от выборки, алгоритма, объема обучающей выборки. Итак, пусть для $i \neq j$ $P_n^A(i, j)$ – *условная вероятность ошибочной классификации* (УОК) нового (не входящего в обучающую выборку) наблюдения из i-го класса в j-й при данной обучающей выборке объема $n = (n_1, \ldots, n_k)$ и алгоритме A. Пусть $\mathbf{E}$ – символ математического ожидания по обучающим выборкам объема n, тогда $\mathbf{E}P_n^A$ называют *ожидаемой ошибкой классификации* (ООК) алгоритма на выборке объема n. Естественно также ввести предел ООК при росте числа наблюдений: $P_\infty^A = \lim\limits_{n \to \infty} \mathbf{E}P_n^A$. P_∞^A называют *асимптотической ожидаемой ошибкой классификации* (АОК).

Часто оказывается, что при $n \to \infty$ и УОК сходится по вероятности к неслучайному пределу. В этом случае этот предел совпадает с P_∞^A. Тем самым пропуск второй буквы «О» в сокращении АОК оправдан. Обычно ООК больше АОК, и отношение

$$\varkappa_n^A = \mathbf{E} P_n^A / P_\infty^A \tag{2.5}$$

характеризует относительное качество обучения алгоритма на выборке объема n. Это очень важный показатель, широко используемый в теоретических и прикладных исследованиях, ввел его Ш. Ю. Раудис, внесший весомый вклад в изучение свойств алгоритмов в условиях дефицита выборочной информации.

Для того чтобы проиллюстрировать масштаб возникающих проблем, в табл. 2.1 приведены значения $\varkappa_{2n}$ для одного из основных алгоритмов дискриминантного анализа — линейной дискриминантной функции, используемой, когда $k = 2$, распределения в классах предполагаются многомерными нормальными (см. п. 2.1.1). При этом предполагается также, что в обучающей выборке имеется ровно по n наблюдений из каждого класса. Параметр d в таблице — это корень из расстояния Махаланобиса между классами (см. п. 1.2.4).

2.1.3. Функции потерь. Ограничимся случаем двух классов. Пусть y определено, как в (2.1); $\widehat{y}_n^A(X)$ — решение, принятое в точке X при использовании решающего правила, построенного с помощью алгоритма A на данной выборке объема n; $q(u, v)$ — функция потерь, такая, что $q(u, u) = 0$ и $q(u, v) > 0$ при $u \neq v$.

Величину

$$\mathbf{E}\{q(y, \widehat{y}_n^A(X)) \mid W_n\} = Q(A, W_n), \tag{2.6}$$

где математическое ожидание берется по всем возможным парам (X, y) при выбранной модели данных, естественно назвать *функцией средних потерь алгоритма A при обучающей выборке W_n*. Если $q(u, v) = (u - v)^2$ и $y \in \{1, 2\}$, то $Q(A, W_n)$ — это средняя ошибка классификации правила, построенного с помощью A на выборка W_n. Взяв математическое ожидание по всем обучающим выборкам объема n, получаем

$$Q_n(A) = \mathbf{E} Q(A, W_n) — \tag{2.7}$$

функцию ожидаемых потерь алгоритма A на обучающей выборке объема n.

Таблица 2.1 [132]

	$d=2,56$ $P_\infty=0,10$				$d=4,65$ $P_\infty=0,01$				$d=6,18$ $P_\infty=0,001$			
	$p=3$	$p=5$	$p=8$	$p=20$	$p=3$	$p=5$	$p=8$	$p=20$	$p=3$	$p=5$	$p=8$	$p=20$
$n=0,6p$	3,26	3,64	3,38	3,44	20,61	25,68	20,93	21,08	162,58	212,62	153,81	147,83
$n=p$	2,22	2,21	2,19	2,16	8,00	7,37	6,84	6,22	39,96	31,86	26,32	20,55
$n=2p$	1,50	1,51	1,51	1,51	2,74	2,66	2,59	2,50	6,16	5,42	4,94	4,47
$n=5p$	1,17	1,18	1,18	1,19	1,45	1,45	1,45	1,44	1,89	1,86	1,83	1,80
$n=10p$	1,08	1,09	1,09	1,09	1,20	1,20	1,20	1,20	1,35	1,35	1,35	1,34
$n=20p$	1,04	1,04	1,04	1,05	1,09	1,10	1,10	1,10	1,16	1,16	1,16	1,16
$n=50p$	1,02	1,02	1,02	1,02	1,04	1,04	1,04	1,04	1,06	1,06	1,07	1,07

Поскольку теоретические распределения не всегда известны, в качестве оценки $Q(A, W_n)$ рассматривают

$$Q_{\text{эмп}}(A, W_n) = \Sigma q(y_i, \widehat{y_n^A}(X_i))/n \qquad (2.8)$$

и называют его *эмпирической функцией средних потерь алгоритма A на выборке* W_n.

2.2. Методы изучения алгоритмов ДА

2.2.1. Базовые асимптотики. В математической статистике принято доверительные интервалы и дисперсии оценок для конечного объема выборки n приближенно представлять через асимптотические (при $n \to \infty$) распределения соответствующих оценок [11, § 8.4]. Это позволяет не только получать хорошие приближения, но и делает теорию более наглядной. Аналогичный прием используется и в дискриминантном анализе. Однако здесь в зависимости от особенностей реальной задачи используются разные асимптотики. Остановимся на этом вопросе более подробно.

Каждый естествоиспытатель знает, что чем больше наблюдений (n) каких-либо статистических объектов он имеет, тем на большее число вопросов относительно характеристик этих объектов он может ответить. Другими словами, чем больше информации, тем более сложная математическая модель может рассматриваться. Если ввести некоторый показатель «сложности» математической модели (C), то различные варианты связи между C и объемом выборки могут быть представлены графически (рис. 2.1).

Горизонтальная прямая (1) на рисунке отвечает традиционной асимптотике математической статистики, используемой главным образом при оценке параметров распределений. Здесь сложность модели — число параметров — фиксирована, а объем выборки n растет [11, § 8.1]. Для регрессионных задач, в которых с ростом числа наблюдений увеличивается число параметров, используемых для описания регрессионной кривой [12, § 6.3, 10.2], более характерна кривая (2), у которой рост C пропорционален n^{α}, где $\alpha < 1$. В задачах статистической классификации широкое распространение получила *асимптотика Колмогорова — Деева* [12, п. 4.3.3], в которой сложность модели — размерность пространства наблюдений p — растет прямо пропорционально числу наблюдений (кривая 3). Кривые на рис. 2.1 пересекаются в одной точке с абсциссой n_0. Эта точка соответствует *вероятностной модели*, которую строит исследо-

ватель, имеющий данное число наблюдений, параметров и т. п. Для всех кривых на рис. 2.1 при $n = n_0$ вероятностная модель одна и та же, однако асимптотические (при $n \to \infty$) свойства изучаемого метода классификации существенно зависят от того, какую модель усложнения вероятностной модели с ростом n, или, другими словами, какую *модель развития* вероятностной модели выберет исследователь. В [26] предлагается объединение вероятностной модели и мо-

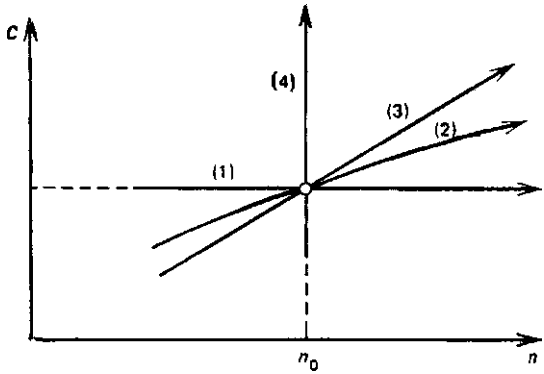


Рис. 2.1. Основные асимптотики: (1) — p фиксировано, $n \to \infty$ (традиционная); (2) — $p = n^{\alpha}$ $(\alpha < 1)$; $n \to \infty$; (3) — p, $n \to \infty$, $p/n \to \lambda < \infty$ (Колмогорова — Деева); (4) — $p \to \infty$, n фиксировано

дели ее развития называть *статистической моделью*. Правда, само понятие модели развития в [26] трактуется шире, чем здесь.

Рассмотрим асимптотики, используемые в теории статистической классификации. В *традиционной асимптотике* математическая модель не меняется, только объем выборки $n \to \infty$. Эксперименты с моделированием выборок из нормальных распределений показывают, что асимптотические (по $n \to \infty$) разложения для ошибок классификации, полученные в традиционной асимптотике [132, 135], близки к результатам моделирования только при $n_i \gg p$.

Используется также *схема серий выборок с моделью растущей* (одновременно с ростом объема выборки) *размерности пространства наблюдений* (асимптотика Колмогорова — Деева). Рассматривается последовательность задач классификации (по параметру $m \to \infty$), при переходе от одной задачи к другой одновременно растут $p = p(m)$ — размерность пространства наблюдений и $n_j = n_j(m)$ — число наб-

людений в обучающей выборке из j-го класса, $j = 1, \ldots, k$. В асимптотике предполагается, что

$$n_j,\ p \to \infty;\quad p/n_j \to \lambda_j < \infty\ (m \to \infty). \tag{2.9}$$

Для изучения статистической задачи классификации эта асимптотика была предложена в 1968 г. А. Н. Колмогоровым, первым обратившим внимание на то, что классификация в задаче Фишера существенно конечномерна. А. Д. Деев [55] исследовал задачу Фишера (см. п. 2.3.1). Хорошее совпадение полученных асимптотических формул для асимптотических ошибок классификации с результатами моделирования [135] привлекло к этой асимптотике внимание теоретиков. Несколько раньше при изучении распределений случайных матриц, связанных с физическими задачами, асимптотика растущей размерности использовалась в работе [103].

Асимптотика растущей размерности при фиксированном числе наблюдений [304] пока носит чисто теоретический характер.

2.2.2. Инвариантность и подобие алгоритмов. В дискриминантном анализе при небольшом числе принципов построения правил классификации предложено очень много конкретных алгоритмов. Поэтому весьма настоятельной является задача выделения алгоритмов в чем-то похожих.

Введем необходимые обозначения. Пусть W_n означает выборку (2.1) объема n; F_j — распределение X, принадлежащих j-му классу; (X, y) — новое наблюдение, не зависящее от W_n; A — алгоритм, $A(W)$ — правило классификации, построенное с помощью алгоритма A на выборке W; $y^{A(W)}(X)$ — результат применения к наблюдению X правила классификации $A(W)$; $G = \{g(X)\}$ — группа преобразований пространства R^p; $g(W_n) = \{(g(X_i), y_i),\ i = 1, \ldots, n\}$. Будем говорить, что алгоритм A *инвариантен относительно* G, если

$$y^{A(W)}(X) = y^{A(g(W))}(g(X)). \tag{2.10}$$

Два алгоритма A и B будем называть *асимптотически* (в традиционной асимптотике) *подобными* для семейства распределений M, если для любого $\varepsilon > 0$ и любых $F_j \in M$ $(j = 1, \ldots, k)$, таких, что $F_j \neq F_{j'}$ $(j \neq j')$, найдется n_0, такое, что для $n > n_0$

$$P\{y^{A(W_n)}(X) = y^{B(W_n)}(X)\} \geqslant 1 - \varepsilon. \tag{2.11}$$

Для асимптотики Колмогорова — Деева в вышеприведенном определении слова «любых $F_j \in M$» надо заменить

на «любой последовательности (по m) $F_j \in M$, удовлетворяющей условиям асимптотики». Причем в условия асимптотики должно обязательно включаться требование стремления расстояний между распределениями к конечным и отличным от нуля пределам.

Пусть $\gamma(c)$ — правило классификации по отношению правдоподобия (1.2), алгоритм A будем называть *состоятельным в традиционной асимптотике над* M, если для любого $\varepsilon > 0$ и любых $F_j \in M$ $(j = 1, 2)$, $F_1 \neq F_2$, найдется n_0, такое, что для $n > n_0$

$$P\left\{y^{A(W_n)}(X) = y^{\gamma(c)}(X)\right\} \geqslant 1 - \varepsilon. \tag{2.12}$$

Для асимптотики Колмогорова — Деева понятие состоятельности алгоритма не вводится, поскольку в ней даже для многомерных нормальных распределений

$$\sup_x \left| \frac{f(X, \widehat{\boldsymbol{\theta}}_2)}{f(X, \widehat{\boldsymbol{\theta}}_1)} - \frac{f(X, \boldsymbol{\theta}_2)}{f(X, \boldsymbol{\theta}_1)} \right| \not\to 0. \tag{2.13}$$

Приведем несколько примеров использования введенных выше понятий.

В п. 1.2.3 описан класс алгоритмов построения древообразных правил классификации в условиях полного знания распределений в классах. Если в формулах (1.48), (1.50) заменить функцию потерь Q и вероятности событий на соответствующие оценки типа (2.8) и частоты, оцененные по выборке W_n, то получим класс древообразных классификаторов. Древообразные классификаторы, очевидно, инвариантны относительно произвольных монотонных преобразований координат и не инвариантны относительно группы общих линейных преобразований R^p. Более того, если в традиционной асимптотике с ростом объема выборки n $\varepsilon_i \to 0$, но так, что $\varepsilon_i n \to \infty$, то для достаточно гладких распределений $F_1 \neq F_2$ древообразные классификаторы асимптотически минимизируют используемую функцию потерь и, в частности, при выборе в (1.47) $q(u, v) = (u - v)^2$ асимптотически подобны байесовскому классификатору и, следовательно, $\gamma(\pi_1/\pi_2)$ состоятельны.

Подстановочный алгоритм, использующий классифицируемое наблюдение (X, y) при оценке неизвестных параметров распределений в модели Фишера. Обозначим $\widehat{M}_{j1}$, $\widehat{M}_{j2}$, $\widehat{\boldsymbol{\Sigma}}_j$ оценки параметров модели M_1, M_2, $\boldsymbol{\Sigma}$ по объединенной выборке $\{W_n, (X, y)\}$, где x — наблюдение, которое предстоит классифицировать, и y положено равным j.

Тогда критерий отношения правдоподобия может быть представлен в виде отношения

$$\widehat{\gamma}=\frac{L\left(W_n,\ X,\ \widehat{M}_{21},\ \widehat{M}_{22},\ \widehat{\Sigma}_2\right)}{L\left(W_n,\ X,\ \widehat{M}_{11},\ \widehat{M}_{12},\ \widehat{\Sigma}_1\right)}.$$

Нетрудно найти новые оценки (в них $\overline{X}_1$, $\overline{X}_2$, $\mathbf{S}$ — традиционные оценки (2.2) и (2.3)):

$$\widehat{M}_{11}=(n_1\overline{X}_1+X)/(n_1+1)=\overline{X}_1+(X-\overline{X}_1)/(n_1+1);$$

$$\widehat{M}_{12}=\overline{X}_2;$$

$$\widehat{\Sigma}_1=\left[(n-2)\,\mathbf{S}+\frac{n_1}{n_1+1}(X-\overline{X}_1)(X-\overline{X}_1)'\right]\Big/(n-1);$$

$$\widehat{M}_{21}=\overline{X}_1;$$

$$\widehat{M}_{22}=\overline{X}_2+(X-\overline{X}_2)/(n_2+1);$$

$$\widehat{\Sigma}_2=\left[(n-2)\,\mathbf{S}+\frac{n_2}{n_2+1}(X-\overline{X}_2)(X-\overline{X}_2)'\right]\Big/(n-1).$$

Итак, в силу очевидных сокращений

$$\widehat{\gamma}=\left(\frac{|\widehat{\Sigma}_1|}{|\widehat{\Sigma}_2|}\right)^{(n+1)/2}.$$

Из-за того, что матрицы добавок к $\mathbf{S}$ в определении $\widehat{\Sigma}_1$ и $\widehat{\Sigma}_2$ имеют ранг 1, формула для γ может быть упрощена:

$$\widehat{\gamma}=\left(\frac{1+\dfrac{n_1}{(n_1+1)(n-2)}(X-\overline{X}_1)'\,\mathbf{S}^{-1}(X-\overline{X}_1)}{1+\dfrac{n_2}{(n_2+1)(n-2)}(X-\overline{X}_2)'\,\mathbf{S}^{-1}(X-\overline{X}_2)}\right)^{(n+1)/2}\approx$$

$$\approx\exp\left\{-\frac{1}{2}\cdot\frac{n+1}{n-2}\cdot\left[\frac{n_2}{n_2+1}\cdot(X-\overline{X}_2)'\,\mathbf{S}^{-1}(X-\overline{X}_2)-\right.\right.$$

$$\left.\left.-\frac{n_1}{n_1+1}\cdot(X-\overline{X}_1)'\,\mathbf{S}^{-1}(X-\overline{X}_1)\right]\right\}.$$

Можно доказать, что подстановочные алгоритмы с использованием пары (X, y) и без нее при оценке параметров асимптотически подобны и в традиционной асимптотике, и в асимптотике растущей размерности.

2.2.3. Методы выработки рекомендаций. Опыт показывает, что статистические феномены, с которыми сталкивается ис-

следователь, работающий в области классификации, сложны и не всегда предсказуемы. Законченных теоретических результатов, которые могли бы выполнять роль ориентиров для практики, по сравнению с другими областями прикладной статистики, здесь заметно меньше. В связи с этим заслуживает специального обсуждения сложившаяся практика осмысления происходящего и выработки соответствующих рекомендаций.

Возникшее у исследователя первоначальное наблюдение или предположение-гипотеза (об источнике ошибок, путях улучшения используемого правила классификации и т. п.) проверяется методом статистического моделирования с известной теоретической моделью на предмет существования. Когда существование феномена установлено, проводится теоретическое исследование, чтобы на простейших математических моделях как в обычной асимптотике, так и в асимптотике растущей размерности Колмогорова — Деева понять его действующие механизмы и дать их асимптотическое количественное описание. На этой стадии обычно не удается получить ни оценок скорости достижения асимптотических утверждений в изучаемых крайне идеализированных моделях, ни границ их применимости. Поэтому необходимо повторное применение метода статистического моделирования, но на этот раз уже с учетом качественного и количественного понимания, достигнутого на простейших моделях. Окончательная проверка полученных рекомендаций проводится на реальных данных.

Состояние теоретических исследований в различных задачах статистической классификации описывается в последующих параграфах настоящей главы. Однако, прежде чем переходить к ним, необходимо отметить, что главное неизвестное любой статистической модели — это состояние природы. Поэтому для конкретных областей применения классификации, с точки зрения прикладной статистики, задачей номер один является накопление примеров эффективного применения конкретных моделей, методов, упрощающих предположений.

2.3. Подстановочные алгоритмы в асимптотике растущей размерности

Как уже сказано в п. 2.1.1, подстановочным (plug-in) алгоритмом называют метод построения правила классификации, при котором неизвестные в отношении правдоподобия параметры распределений Θ заменяют их оценками максималь-

ного правдоподобия $\widehat{\Theta}$. При минимальных требованиях к плотности распределений подстановочные алгоритмы в традиционной асимптотике асимптотически подобны и $\gamma(c)$ состоятельны. Как следует из формулы (2.13), положение в случае асимптотики растущей размерности сложнее. Здесь уже многое зависит от того, как оцениваются параметры и насколько эффективно используются упрощающие предположения.

2.3.1. Модель Фишера в асимптотике (2.9). Базовое предположение (2.9) дополним условием, что

$$J_p = (M_2 - M_1)' \Sigma^{-1} (M_2 - M_1) \to J < \infty, \tag{2.14}$$

т. е. что расстояние Махаланобиса между распределениями стремится к конечному пределу.

Рассмотрим сначала случай, когда Σ известно (см. п. 1.1.2). Согласно (1.12) подстановочное правило классификации имеет вид:

$$h(X) = (X - (\overline{X}_1 + \overline{X}_2)/2)' \Sigma^{-1} (\overline{X}_2 - \overline{X}_1) \gtrless c, \tag{2.15}$$

где $\overline{X}_1$, $\overline{X}_2$ — обычные выборочные средние для обучающих выборок из первой и второй совокупностей. Предположим для определенности, что X извлечено из первой совокупности, и найдем условную вероятность ошибки классификации по правилу (2.15) при фиксированной обучающей выборке

$$P\{H_2 \mid H_1, W_n\} = 1 - \Phi((c - a_1)/\sigma), \tag{2.16}$$

где

$$a_1 = (M_1 - (\overline{X}_1 + \overline{X}_2)/2)' \Sigma^{-1} (\overline{X}_2 - \overline{X}_1), \tag{2.17}$$

$$\sigma^2 = \mathbf{D}[X' \Sigma^{-1} (\overline{X}_2 - \overline{X}_1) \mid H_1, W_n] = (\overline{X}_2 - \overline{X}_1)' \Sigma^{-1} (\overline{X}_2 - \overline{X}_1). \tag{2.18}$$

Аналогично

$$P\{H_1 \mid H_2, W_n\} = \Phi((c - a_2)/\sigma), \tag{2.16'}$$

где

$$a_2 = (M_2 - (\overline{X}_1 + \overline{X}_2)/2)' \Sigma^{-1} (\overline{X}_2 - \overline{X}_1). \tag{2.19}$$

В предположениях (2.9), (2.14) с ростом объема обучающей выборки a_1, a_2, σ^2 сближаются со своими математическими ожиданиями и стремятся соответственно к пределам

$$a_1 \to -J/2 + (-\lambda_1 + \lambda_2)/2, \tag{2.20}$$

$$a_2 \to \quad J/2 + (-\lambda_1 + \lambda_2)/2, \tag{2.21}$$

$$\sigma^2 \to \quad J + \lambda_1 + \lambda_2. \tag{2.22}$$

Из (2.20) — (2.22) видно, что асимптотическое значение α-минимаксной ошибки классификации достигается при равных асимптотических ошибках первого и второго рода, т. е. при $c \to (\lambda_2 - \lambda_1)/2$, и

$$\alpha \to \Phi\left(-J/2\sqrt{J + \lambda_1 + \lambda_2}\right). \tag{2.23}$$

В проведенном выше рассуждении сразу от условной ошибки классификации перешли к асимптотической ошибке, не вычисляя в качестве промежуточного этапа ожидаемую ошибку классификации.

Общая модель с матрицей Σ, оцениваемой по выборочным данным, была изучена А. Д. Деевым [55]. В предположении, что $\lambda_1^{-1} + \lambda_2^{-1} > 1$, он показал, что для подстановочного правила минимаксная ошибка классификации

$$\alpha \to \Phi\left(-J\,(1 - \lambda_1\lambda_2/(\lambda_1 + \lambda_2))^{1/2}/2\sqrt{J + \lambda_1 + \lambda_2}\right). \tag{2.24}$$

Как видно из сравнения формул (2.23) и (2.24), цена (в терминах α), которую приходится платить за $p\,(p+1)/2$ неизвестных параметров общей ковариационной матрицы, достаточно высока. Как уже сказано в п. 2.2.1, формулы Деева дают хорошую аппроксимацию даже при умеренных объемах обучающих выборок. В этом можно убедиться непосредственно, сравнив данные табл. 2.1 и 2.2. В табл. 2.2 приведены асимптотические значения $\varkappa_{2n} = EP_{2n}/P_\infty$ для линейной дискриминантной функции, полученные по формуле (2.23), когда матрица известна, и (2.24) — в общем случае при $\pi_1 = \pi_2 = 0{,}5$.

Таблица 2.2

	Матрица Σ известна			Матрица Σ неизвестна		
	$d=2{,}56$ $P_\infty=0{,}1$	$d=4{,}65$ $P_\infty=0{,}01$	$d=6{,}18$ $P_\infty=0{,}001$	$d=2{,}56$ $P_\infty=0{,}1$	$d=4{,}65$ $P_\infty=0{,}01$	$d=6{,}18$ $P_\infty=0{,}001$
$n=0{,}6p$	1,49	1,52	1,52	3,35	18,85	113,14
$n=p$	1,31	1,30	1,30	2,10	5,78	16,59
$n=2p$	1,17	1,15	1,14	1,51	2,45	4,12
$n=5p$	1,07	1,06	1,06	1,19	1,44	1,77
$n=10p$	1,04	1,03	1,03	1,09	1,20	1,30
$n=20p$	1,02	1,02	1,01	1,05	1,10	1,16
$n=50p$	1,01	1,01	1,00	1,02	1,04	1,06

2.3.2. Распределения с независимыми блоками. Эти распределения введены в п. 1.1.5. Они служат простейшей мо-

делью негауссовских распределений. Добавим к базовым предположениям (2.9) предположения, что размерность векторов $X^{(j)}$ и $\Theta^{(j)}$ в блоках ограничена

$$p_j,\ m_j < c < \infty; \tag{2.25}$$

что значения соответствующих параметров в классифицируемых распределениях сближаются друг с другом:

$$\left|\theta_2^{(j)} - \theta_1^{(j)}\right| < c/\sqrt{n}, \tag{2.26}$$

и что суммарное расстояние между распределениями стремится к конечному пределу

$$\Sigma J_j \to J < c < \infty, \tag{2.27}$$

где $J_j = \sum\limits_{t,s} (\theta_2^{(t)} - \theta_1^{(t)})\, i_{t,s}\,(\Theta_1)(\theta_2^{(s)} - \theta_1^{(s)})$ и суммирование проводится по всем t, s, принадлежащим j-му блоку; $\|i_{t,s}(\Theta)\| = \left\| \int \frac{\partial \ln f(X, \Theta)}{\partial \theta^{(t)}} \cdot \frac{\partial \ln f(X, \Theta)}{\partial \theta^{(s)}} f(X, \Theta)\, \mu(dX) \right\|$ — информационная матрица Фишера [12, § 8.2—8.3]. При выполнении условий (2.9), (2.25) — (2.27) и некоторых дополнительных условий регулярности в [109] показано, что для подстановочного алгоритма справедлива формула (2.23). Более того, если в j-м блоке имеются m_j различающих и l_j неизвестных, но общих обоим распределениям параметров, причем $l_j < c < \infty$, и одна и та же оценка общих параметров подставляется в обе плотности, то (2.23) также имеет место. Другими словами, $O(p)$ общих параметров не ухудшают асимптотические свойства подстановочного алгоритма. Можно надеяться, что и в задаче Фишера в случае, когда Σ зависит только от cp параметров и при оценке этот факт учитывается, (2.23) также будет справедлива. Эту гипотезу удалось доказать в случае древообразных распределений.

2.3.3. Модель Фишера в случае древообразных распределений. Если при древообразных (ДСЗ) распределениях с известной структурой зависимостей оценку Σ проводить не по общей схеме, а с учетом структуры зависимостей, как указано в [12, п. 4.2.3], то согласно [77, 78] асимптотическая минимаксная ошибка модернизированного классификатора будет не (2.24), а (2.23), т. е. существенно меньше. Более того, известно, что при минимальных дополнительных предположениях древообразная структура зависимостей восстанавливается в асимптотике Колмогорова — Деева с точностью до несущественных связей с вероятностью 1 [12, п. 4.3.3 и 4.3.4].

2.3.4. Оцифровка градаций качественных переменных. Если в исследовании встречаются качественные переменные,

то для применения к ним общих линейных моделей дискриминантного анализа их градациям часто приписывают численные значения-метки и далее работают с этими оцифрованными переменными как с обычными числами. При этом используются две стратегии: первая (универсальная) состоит в том, что каждая градация качественной переменной выделяется в новую двоичную переменную, принимающую два значения: 0, если градация не осуществилась, и 1, если осуществилась [11, п. 10.2.4]; вторая стратегия применяется тогда, когда качественные градации можно рассмат-

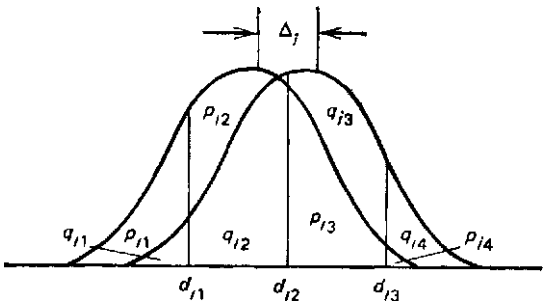


Рис. 2.2. Границы квантования и плотности распределений в задаче об оцифровке качественных переменных

ривать как результат квантования некоторой непрерывной случайной величины (ее математическая техника описана ниже). Наша ближайшая цель — сравнить на простейшей математической модели эффективность этих подходов в асимптотике растущей размерности.

Математическая модель: рассматриваются два класса с независимыми переменными в каждом из классов. Пусть p_{ij} (q_{ij}) — вероятность того, что i-я переменная в первом классе (соответственно во втором) принимает свое j-е значение, $j = 1, \ldots, k_i < c < \infty$, и $\widehat{p_{ij}}$ ($\widehat{q_{ij}}$) — оценки этих вероятностей по частотам на основании обучающих выборок. Предположим, что существует такая известная функция $G(z)$ с непрерывной первой производной; такие числа Δ_i — расстояния между функциями G для первого и второго классов; границы квантования d_{ij}, $i = 1, \ldots, p$; $j = 0, 1, \ldots, k_i$ — $-\infty = d_{i0} < d_{i1} < \ldots < d_{ik_i} = \infty$, такие, что

$$p_{ij} = G(d_{ij} + \Delta_i/2) - G(d_{ij-1} + \Delta_i/2); \qquad (2.28)$$

$$q_{ij} = G(d_{ij} - \Delta_i/2) - G(d_{ij-1} - \Delta_i/2). \qquad (2.29)$$

Это наглядно показано на рис. 2.2.

Пусть далее выполняются следующие асимптотические (в асимптотике растущей размерности) предположения:

$$k_i < c_1 < \infty; \tag{2.30}$$

$$p,\ k,\ n_1,\ n_2 \to \infty;\quad p/n_1 \to \lambda_1,\ p/n_2 \to \lambda_2;$$

$$\Sigma\,(k_i - 1)/n_1 \to \lambda_3,\ \Sigma\,(k_i - 1)/n_2 \to \lambda_4; \tag{2.31}$$

$$p_{ij} - q_{ij} \geqslant c_2 > 0; \tag{2.32}$$

$$\Delta_{ij} = p_{ij} - q_{ij} = \delta_{ij}/\min\,(n_1,\ n_2),\ 0 < c_3 \leqslant |\,\delta_{ij}\,| \leqslant c_4 < \infty; \tag{2.33}$$

$$J_m = \sum_i \sum_j \Delta_{ij}^2/(0{,}5\,(p_{ij} + q_{ij})) \to J < \infty, \tag{2.34}$$

Подготовительные вычисления по второй схеме $r_{ij} = (\widehat{p_{ij}} + \widehat{q_{ij}})/2$, $R_{ij} = \sum\limits_{l \leqslant j} r_{il}$, $b_{ij} = \dot{G}\,(R_{ij}) - \dot{G}\,(R_{ij-1})$, $a_{ij} = b_{ij}/r_{ij}$, $\widehat{\Delta}_i = \sum\limits_j a_{ij}\,(\widehat{p_{ij}} - \widehat{q_{ij}})\,/\sum\limits_j r_{ij}\,a_{ij}^2$.

В данных предположениях: при оцифровке по первой схеме, когда градации оцифровываются независимо друг от друга так, что j-й градации i-й переменной приписывается значение $z_{ij} = \ln\,(\widehat{p_{ij}}/\widehat{q_{ij}})$ и классификация проводится по правилу $\Sigma z_i \gtrless c$, где порог c подбирается из условия минимизации максимальной вероятности ошибки α,

$$\alpha \to \Phi\left(-J/2\sqrt{J + \lambda_3 + \lambda_4}\right); \tag{2.35}$$

при оцифровке по второй схеме, когда j-й градации i-й переменной приписывается значение $z_{ij} = a_{ij}\widehat{\Delta}_i$ и классификация проводится по тому же правилу, выполняется соотношение (2.23). Формулы (2.35) и (2.23) совпадают только в случае, когда у переменных имеются всего по две градации $k_i = 2$. Уже при $k_i = 3$ формула (2.35) дает заметно большую ошибку. Таким образом, независимой оцифровки градаций признаков следует избегать.

2.4. Статистическая регуляризация оценки обратной ковариационной матрицы в линейной дискриминантной функции для модели Фишера

2.4.1. Качественный анализ трудностей линейного дискриминантного анализа в асимптотике растущей размерности. Как показано в п. 2.3.1, замена неизвестной обратной ковариационной матрицы $\mathbf{\Sigma}^{-1}$ ее оценкой $\mathbf{S}^{-1}$ в общем случае

приводит к заметному росту ООК. Это отчасти можно объяснить плохой обусловленностью матрицы $\mathbf{S}$ при $p \sim n$ и тем, что оценка $\mathbf{S}^{-1}$ не является состоятельной в асимптотике растущей размерности, так как

$$\mathbf{E}S^{-1} = \Sigma^{-1}\left(1 - \frac{p}{n}\right)^{-1} + \mathbf{O}\left(\frac{1}{n}\right),$$

где $|\Sigma| > 0$, $p < n$, симметричная $(p \times p)$-матрица $\mathbf{O}$ имеет максимальное собственное число $O\left(\frac{1}{n}\right)$. Для того чтобы понять, в чем дело, зафиксировав обучающую выборку, попытаемся построить наилучшее при данной выборке решающее правило, а затем сравним его с правилом, получаемым при использовании подстановочного алгоритма. При этом оптимальное для УОК правило выведем при использовании дополнительной информации, которой нельзя воспользоваться в обычной практике. Тем не менее сравнение двух правил покажет направления для возможного улучшения подстановочного алгоритма.

Произведем два последовательных преобразования пространства наблюдений: линейное, превращающее обычную ковариационную матрицу в единичную

$$Y = \Sigma^{-1/2}(X - (M_1 + M_2)/2),$$

и ортогональное, ориентирующее координатные оси вдоль направлений собственных векторов выборочной ковариационной матрицы в пространстве Y:

$$\mathbf{S}_Y = \sum_{j=1}^{2} \sum_{i\ y_i = j} (Y_i - \overline{Y}_j)(Y_i - \overline{Y}_j)'/(n-2),$$

$Z = \mathbf{C}_Y' Y$, где $\mathbf{C}_Y$ $(p \times p)$-матрица, составленная из собственных векторов матрицы $\mathbf{S}_Y$. В пространстве Z выборочная ковариационная матрица диагональна и дискриминантная функция имеет простой вид

$$h(Z) = \sum_{i=1}^{p} \delta_i^{-1}(z^{(i)} - (\overline{z}_1^{(i)} + \overline{z}_2^{(i)})/2)(\overline{z}_2^{(i)} - \overline{z}_1^{(i)}). \qquad (2.36)$$

Рассмотрим теперь функцию $\tilde{h}(Z)$ вида

$$\tilde{h}(Z) = \sum_{i=1}^{p} \alpha_i (z^{(i)} - (\overline{z}_1^{(i)} + \overline{z}_2^{(i)})/2)(\overline{z}_2^{(i)} - \overline{z}_1^{(i)}), \qquad (2.37)$$

где α_i — постоянные, подобранные так, чтобы

$$d = |\mathbf{E}(\tilde{h}(Z) \mid H_2,\ \mathbf{S}_Y) -$$

$$- \mathbf{E}(\tilde{h}(Z) \mid H_1,\ \mathbf{S}_Y)| / \sqrt{\mathbf{D}(\tilde{h}(Z) \mid H_1,\ \mathbf{S}_Y)},$$

а следовательно, и УОК были оптимальны. Находим

$$\alpha_i \sim \frac{d^{(i)}}{d^{(i)} + \xi_2^{(i)} + \xi_1^{(i)}}. \qquad (2.38)$$

здесь $d^{(i)} = \mathbf{E}(z^{(i)} | H_2, \mathbf{S}_Y) - \mathbf{E}(z^{(i)} | H_1, \mathbf{S}_Y)$; $\xi_j^{(i)} = \bar{z}_j^{(i)} + (-1)^{j+1} d^{(i)}/2 \in N(0, n_j^{-1})$ $(j = 1, 2; i = 1, \ldots, p)$ и независимы между собой. В рассуждении использовано то обстоятельство, что $(\bar{Y}_1, \bar{Y}_2)$ и $\mathbf{S}_Y$ независимы между собой и ковариационная матрица $\mathbf{Y}$ единичная.

Сравним теперь формулы (2.36)—(2.38):

1) в традиционной асимптотике при $d^{(i)} \neq 0$ $\alpha_i \to 1$, аналогично $\delta_i \to 1$, поэтому обычный линейный дискриминантный анализ и алгоритм оптимизации УОК асимптотически подобны;
2) теоретически [51, 103, 142] и путем моделирования показано, что в асимптотике растущей размерности δ_i не стремятся к пределу, а имеют предельное распределение с размахом, зависящим от λ_i, $i = 1, 2$; распределение δ_i не зависит от $d^{(i)}$ и $\xi_j^{(i)}$, поэтому взвешивание не оптимально и линейный дискриминантный анализ ведет к большим по сравнению с алгоритмом (2.37) — (2.38) ошибкам (напомним, что последний алгоритм использует информацию об истинных параметрах модели);
3) из-за нормализующего преобразования $X \to Y$ алгоритм евклидова расстояния в пространстве Y, относящий наблюдение к той совокупности, к выборочному центру которой оно ближе, может иметь меньшую ООК по сравнению с линейной дискриминантной функцией;
4) алгоритмы, уменьшающие вклад в дискриминантную функцию экстремальных значений δ_i как источника больших погрешностей и учитывающие при выборе весов в (2.37) величину $d^{(i)}$, могут в асимптотике растущей размерности вести к уменьшению ООК по сравнению с традиционным дискриминантным анализом. Особенно опасны δ_i, близкие к нулю.

2.4.2. Регуляризованные оценки $\mathbf{S}^{-1}$. Специальные меры, направленные на улучшение обусловленности матрицы $\mathbf{S}$ и уменьшение случайных колебаний корней обратной матрицы $\mathbf{S}^{-1}$, принято называть *регуляризацией*. Пусть X —

собственный вектор матрицы $\mathbf{S}$, соответствующий собственному числу δ, т. е.

$$\mathbf{S}X = \delta X. \tag{2.39}$$

Тогда X является собственным вектором матрицы $\mathbf{I}_p + a\mathbf{S}$ $(a > 0)$, соответствующим собственному числу $1 + a\delta$, так как

$$(\mathbf{I}_p + a\mathbf{S})\, X = X + a\delta X = (1 + a\delta)\, X. \tag{2.40}$$

Заменим теперь в линейной дискриминантной функции предыдущего пункта $\mathbf{S}_Y^{-1}$ на $(\mathbf{I}_p + a\mathbf{S}_Y)^{-1}$, тогда в силу сохранения собственных векторов представление (2.36) имеет место, и в нем величины δ_i^{-1} заменяются на $(1 + a\delta_i)^{-1}$. Разброс последних заведомо меньше разброса δ_i^{-1}, они ближе к предельному взвешиванию слагаемых и, следовательно, обеспечивают меньшую ООК, чем (2.36). При $a = 0$ получаем алгоритм евклидова расстояния.

К сожалению, невозможно воспользоваться только что проведенным рассуждением непосредственно, так как исходная матрица $\mathbf{\Sigma}$ неизвестна. Однако на практике регуляризация рассмотренного вида часто применяется к исходной выборочной ковариационной матрице (без предварительного перехода в пространство Y). При этом, так же как в рассмотренном выше случае, направления собственных векторов не меняются, а собственные числа матрицы отодвигаются от нуля. Это так называемые ридж-оценки $\mathbf{S}^{-1}$. В работе [23] теоретически и в [217] путем моделирования показано, что ридж-оценки действительно уменьшают ООК. В [167] подобный результат достигается при замене $\mathbf{S}^{-1}$ на $(\mathbf{S} + a\mathbf{A})^{-1}$, где $\mathbf{A}$ — некоторая симметричная положительно определенная матрица. В частности, в качестве $\mathbf{A}$ можно взять матрицу, составленную из диагональных элементов $\mathbf{S}$.

Другой вид регуляризации, с успехом используемый на практике [148] и называемый оценкой главных компонент (ОГК-оценкой) — это замена $\mathbf{S}^{-1}$ на $\mathbf{S}_\gamma^{-1} = \mathbf{C}\ \mathrm{diag}\ (\delta_1^{-1} V(\delta_1 - \gamma),\ \ldots,\ \delta_p^{-1} V(\delta_p - \gamma))\ \mathbf{C}'$,
где $\mathbf{C}$ — ортогональная $(p \times p)$-матрица, составленная из собственных векторов матрицы $\mathbf{S}$; $(\delta_1, \ldots, \delta_p)$ — собственные числа матрицы $\mathbf{S}$, а $V(u) = 0$ для $u \leqslant 0$ и $V(u) = 1$ для $u > 0$.

Простая геометрическая иллюстрация рассмотренных выше правил дана на рис. 2.3 посредством функций взвешивания собственных значений матрицы $\mathbf{S}$. Пусть $\mathbf{C}$ и $\{\delta_1, \ldots, \delta_p\}$ определены как выше и пусть $U = \mathbf{C}'X$, тогда в тер-

минах U линейная дискриминантная функция представляется в виде

$$h(X)=\widetilde{h}(U)=\sum_{i=1}^{p}\delta_i^{-1}(U-(\overline{U}_1+\overline{U}_2)/2)'(\overline{U}_2-\overline{U}_1). \qquad (2.41)$$

Введем в (2.41) формально в виде сомножителя функцию взвешивания $\eta(\delta_i)$. Это позволяет единообразно представить основные ортогонально-инвариантные методы регуляризации: $а$ — $\eta(\delta)\equiv 1$ для линейной дискриминантной функции Фишера; $б$ — $\eta(\delta)=\begin{cases}0 & \text{для } \delta\leqslant\gamma\\ 1 & \delta>\gamma\end{cases}$ для ОГК-оценки S^{-1}; $в$ — $\eta(\delta)=\delta$ для метода евклидова расстояния; $г$ — $\eta(\delta)=\dfrac{\delta}{\alpha+\delta}$ для ридж-оценок вида $(\alpha I_p+S)^{-1}$.

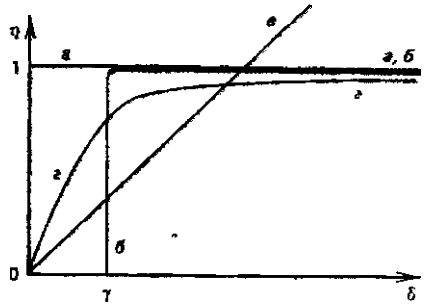


Рис. 2.3. Весовые коэффициенты в различных методах регуляризации S

2.4.3. Обобщенная ридж-оценка В.И. Сердобольского [142, 145]. Представляет собой линейную комбинацию простых ридж-оценок $(I+tS)^{-1}$ с функцией взвешивания $a(t)$, где $a(t)$ — функция ограниченной вариации

$$S_a^{-1}=\int_{t\geqslant 0}(I_p+tS)^{-1}\,da(t). \qquad (2.42)$$

Для того чтобы для заданной функции $a(t)$ при использовании S_a^{-1} вместо S^{-1} в линейной дискримнинантной функции существовало в асимптотике растущей размерности предельное распределенне для УОК, предположения (2.9) должны быть дополнены следующими:

1) обе совокупностн нормальны $N(M_j,\ \Sigma)$, $j=1,\ 2$;

2) собственные числа матриц Σ лежат на отрезке $[c_1, c_2]$, где $c_1>0$ и c_2 от m не зависят;

3) при каждом m сумма $n_1+n_2\geqslant p+4$, $0<\lambda=\lim\limits_{m\to\infty} p/(n_1+n_2)<1$.

Введем функцию распределения неслучайных собственных значений матрицы Σ: $F_m(u)=p^{-1}\sum\limits_{i:\delta_i\leqslant u,\ i=1,\ \ldots,\ p}1$.

Обозначим $\mu=M_2-M_1$, и пусть ниже $\mu^{(i)}$ означают компоненты вектора μ в системе координат, в которой

матрица Σ диагональна. Введем функцию $R_m(u) = \sum_{i\ \delta_i \leqslant u,\ i=1,\ldots,p} \mu^{(i)2}/\delta_i$;

4) при $u \geqslant 0$ существуют пределы

$$F(u) = \lim_{m\to\infty} F_m(u),\ R(u) = \lim_{m\to\infty} R_m(u);$$

5) выборки из совокупностей независимы. Матрица S вычисляется обычным образом согласно (2.3).

В сделанных предположениях существуют пределы в среднем квадратическом:

$$h(z) = \text{l. i. m. } p^{-1}\,\mathrm{Sp}\,(\mathbf{I}_p - z\mathbf{S})^{-1} = \int (1 - zs(z)u)^{-1}\,\mathrm{d}F(u), \tag{2.43}$$

где $s(z) = 1 - \lambda + \lambda h(z)$;

$$b(z) = \text{l. i. m. } \mu'(\mathbf{I}_p - z\mathbf{S})^{-1}\mu = \int (1 - zs(z)u)^{-1} u\,\mathrm{d}R(u); \tag{2.44}$$

$$k(z) = \lim_{m\to\infty} \text{l. i. m. } (\overline{X}_2 - \overline{X}_1)'(\mathbf{I} - z\mathbf{S})^{-1}(\overline{X}_2 - \overline{X}_1) =$$

$$= b(z) + (\lambda_1 + \lambda_2)(h(z) - 1)/(zs(z)). \tag{2.45}$$

Предельная минимаксная ошибка (α) классификации по правилу $h(X) \gtrless \theta_0$, где $\theta_0 = \lim_{m\to\infty} \text{l.i.m. } \theta_{0m} = \frac{1}{2}(\lambda_2 - \lambda_1) \times \int (ts(-t))^{-1}(1 - h(-t))\,\mathrm{d}a(t)$, выражается через них:

$$\alpha = \Phi\left(-\sqrt{G^2/D}/2\right), \tag{2.46}$$

где

$$G = \lim_{m\to\infty} \text{l. i. m. } \mu'\mathbf{S}_a^{-1}\mu = \int b(-t)\,\mathrm{d}a(t); \tag{2.47}$$

$$D = \text{l. i. m. } (\mu'\mathbf{S}_a^{-1}\mathbf{\Sigma}\mathbf{S}_a^{-1}\mu + (n_1^{-1} + n_2^{-1})\,\mathrm{Sp}\,\mathbf{\Sigma}\mathbf{S}_a^{-1}\mathbf{\Sigma}\mathbf{S}_a^{-1}) =$$

$$= \int\int (s(-t)s(-t_1))^{-1}\frac{k(-t) - k(-t_1)}{t_1 - t}\,\mathrm{d}a(t)\,\mathrm{d}a(t_1). \tag{2.48}$$

В [142] доказывается, что функция $a_0(t)$, минимизирующая α, может быть найдена при некоторых дополнительных предположениях в явном виде. Переход от предельных рекомендаций [142] к построению практического алгоритма для конечных выборок является довольно сложной задачей и выполнен в [145]. Соответствующая программа, названная ЭЛДА — экстремальный линейный дискриминантный анализ — хорошо работает начиная с $p \geqslant 5$. Ее сравнение с

алгоритмами Фишер и Парзен (см. § 3.2) на ряде реальных коллекций данных с $p \sim n$, выполненное с помощью пакета СОРРА-2 [125], показало явное преимущество ЭЛДА перед алгоритмом Фишер и заметный выигрыш по сравнению с универсальным алгоритмом Парзен при использовании в последнем стандартных значений параметра сглаживания, предусмотренных в СОРРА-2.

2.5. Отбор переменных

2.5.1. Увеличение ООК малоинформативными признаками. Один из очевидных выводов из формул § 2.3 состоит в том, что включение в прогностическое правило малоинформативных переменных может заметно ухудшить его качество. Рис. 2.4 показывает это наглядно. Каждый признак наряду

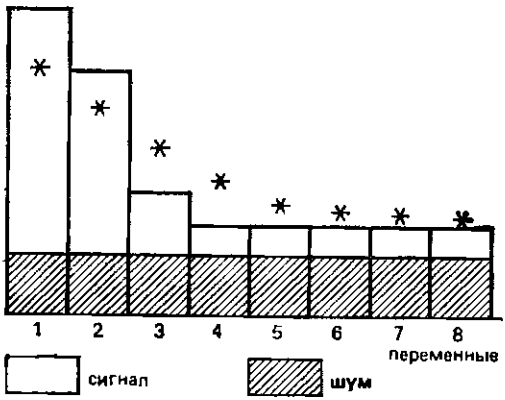


Рис. 2.4. Зависимость отношения сигнал/шум от числа отобранных параметров: ✻ — отношение сигнал/шум для k первых переменных

с положительным вкладом в разделение несет в себе в силу ограниченности выборки и шумовую (случайную) составляющую. Если много малоинформативных признаков, то отношение сигнал/шум значительно лучше для группы высокоинформативных признаков, чем для всей выборки. Тот же вывод подтверждают и числовые данные.

Из анализа данных табл. 2.2 видно, что при известной ковариационной матрице Σ обучаемость подстановочного алгоритма заметно лучше, чем в общем случае, когда Σ неизвестна. Однако и при известном Σ роль отношения p/n

существенна. Поэтому при относительно небольшом объеме выборки малоинформативные признаки в прогностическое правило лучше не включать. Однако заранее информативность признаков обычно не известна и отбор наилучших среди них производится по выборке, но здесь мы сталкиваемся с новым явлением — отбор признаков может заметно ухудшить обучаемость алгоритма.

2.5.2. Влияние выборочных флуктуаций на результаты отбора признаков. Задача формирования наилучшей системы признаков трудна сама по себе как с технической, так и с методологической стороны даже в случае полностью определенных распределений (см. § 1.4). В дискриминантном анализе она усугубляется еще и выборочными флуктуациями. Для представления масштаба возникающей проблемы снова обратимся к модельному примеру. Пусть в модели Фишера с известной единичной ковариационной матрицей

$$F_j = N(M_j,\ \mathbf{I}_p)\ (j = 1,\ 2) \tag{2.49}$$

средние случайны:

$$M_1 = -M_2 \text{ и } M_1 \in N(0,\ \sigma^2 \mathbf{I}_p/4). \tag{2.50}$$

При моделировании сначала получают значения M_1 и M_2, далее моделируются независимые выборки объема n каждая из F_1 и F_2, и по ним с помощью изучаемого алгоритма A строится правило классификации. Поскольку значения M_1 и M_2 известны, нетрудно оценить $P^A_{p,\infty}$, — асимптотическую ошибку классификации, которая, естественно, зависит от M_1 и M_2. Подбирая величину σ^2, можно добиться того, что значение $\mathbf{E}P^A_{p,\infty}$ будет достаточно близко к любому числу $0 < \varepsilon < 0{,}5$.

Пусть A — подстановочный алгоритм, действующий в R^p и порождающий правило вида

$$h(X) = (X - (\overline{X}_1 + \overline{X}_2/2))'\ (\overline{X}_2 - \overline{X}_1) \gtrless c, \tag{2.51}$$

где c подбирается в каждой серии так, чтобы УОК была минимаксной. Пусть далее B — аналогичный подстановочный алгоритм, но с предварительным отбором r признаков из p. При этом отбор переменных проводится по величине модуля разности $|\overline{x}_2^{(i)} - \overline{x}_1^{(i)}|$ так, что переменные с разностью, большей некоторого порога, включаются как «информативные», а с меньшей — нет. В табл. 2.3 показаны три отношения $\varkappa^A = \mathbf{E}P^A_{p,2n}/\mathbf{E}P^A_{p,\infty}$, $\varkappa^B = \mathbf{E}P^B_{p,2n}/\mathbf{E}P^B_{p,\infty}$ и $\gamma = (\varkappa^B - 1)\,/\,(\varkappa^A - 1)$, полученные методом статистического моделирования. Общий вывод, который можно сделать из табл. 2.3, следующий: в рассматриваемой моде-

Таблица 2.3 [133]

			$\rho=0,125$			$\rho=0,25$			$\rho=0,5$		
P_∞	r	n/r	$\varkappa^A$	$\varkappa^B$	γ	$\varkappa^A$	$\varkappa^B$	γ	$\varkappa^A$	$\varkappa^B$	γ
0,10	5	0,5	1,78	2,73	2,2	1,73	2,45	1,9	1,74	1,98	1,3
		1	1,40	2,15	2,9	1,39	1,86	2,19	1,39	1,6	1,5
		2	1,19	1,64	3,4	1,31	1,57	2,8	1,17	1,30	1,8
		5	1,08	1,25	3,3	1,08	1,20	2,6	1,08	1,14	1,7
		10	1,04	1,13	3,3	1,04	1,10	2,5	1,04	1,06	1,6
0,10	8	0,5	1,74	2,67	2,2	1,76	2,41	1,9	1,66	2,00	1,5
		1	1,38	2,05	2,8	1,37	1,82	2,2	1,35	1,56	1,6
		2	1,19	1,59	3,2	1,20	1,50	2,5	1,19	1,33	1,7
		5	1,08	1,28	3,5	1,07	1,20	2,8	1,08	1,13	1,7
		10	1,04	1,15	4,2	1,04	1,10	2,7	1,04	1,07	1,7
0,10	12	0,5	1,73	2,52	2,1	1,58	2,23	2,1	1,66	1,96	1,5
		1	1,35	2,03	2,9	1,35	1,79	2,3	1,35	1,55	1,57
		2	1,18	1,58	3,3	1,17	1,45	2,6	1,17	1,30	1,8
		5	1,07	1,25	3,6	1,07	1,19	2,6	1,07	1,13	1,9
		10	1,04	1,13	3,5	1,04	1,10	2,9	1,04	1,07	1,8
0,10	20	0,5	1,63	2,52	2,4	1,63	1,2	1,9	1,65	1,96	1,5
		1	1,35	2,02	2,9	1,33	1,77	2,4	1,34	1,59	1,7
		2	1,18	1,59	3,4	1 17	1,45	2,6	1,18	1,32	1,8
		5	1,07	1,24	3,4	1,07	1,19	2,7	1,07	1,13	1,8
		10	1,03	1,13	3,8	1,04	1,10	2,7	1,04	1,07	1,9
0,01	5	0,5	2,36	5,94	3,6	2,44	3,96	2,1	2,09	2,73	1,6
		1	1,58	2,97	3,4	1,63	2,57	2,5	1,60	1,93	1,6
		2	1,29	1,83	2,9	1,25	1,66	2,7	1,23	1,38	1,7
		5	1,11	1,30	2,64	1,10	1,22	2,10	1,08	1,13	1,6
		10	1,05	1,13	2,62	1,06	1,13	2,27	1,05	1,08	1,6
0,01	8	0,5	2,21	5,14	3,4	2,19	3,88	2,4	2,07	2,80	1,7
		1	1,51	2,61	3,2	1,48	2,30	2,7	1,46	1,78	1,7
		2	1,25	1,84	3,4	1,23	1,54	2,3	1,25	1,41	1,68
		5	1,09	1,31	3,3	1,08	1,19	2,3	1,08	1,12	1,5
		10	1,04	1,14	3,2	1,04	1,10	2,4	1,04	1,07	1,6
0,01	12	0,5	1,99	4,42	3,4	2,03	3,34	2,3	1,93	2,53	1,6
		1	1,43	2,61	3,7	1,42	2,29	2,8	1,45	1,75	1,7
		2	1,21	1,72	3,4	1 20	1,52	2,5	1,22	1,35	1,6
		5	1,08	1,27	3,4	1,09	1,21	2,4	1,09	3,13	1,5
		10	1,04	1,14	3,7	1,04	1,10	2,3	1,04	1,06	1,5
0,01	20	0,5	1,85	4,25	3,8	1,78	3,20	2,9	1,84	2,53	1,8
		1	1,38	2,43	3,8	1,40	2,01	2,5	1,37	1,64	1,7
		2	1,19	1,70	3,7	1,13	1,38	1,6	1,17	1,29	1,7
		5	1,07	1,28	3,8	1,07	1,19	2,73	1,07	1,11	1,6
		10	1,04	1,12	3,5	1,04	1,09	2,61	1,03	1,05	1,56

ли, когда объем обучающей выборки ограничен и число отобранных признаков в 4—8 раз меньше числа исходных переменных, ожидаемая ошибка алгоритма с отбором признаков по обучающей выборке заметно больше ожидаемой ошибки алгоритма без отбора. Правда, в качестве примера взята модель ситуации, весьма трудной для отбора.

2.5.3. Изучение эффекта отбора признаков в асимптотике растущей размерности. Основное добавление к предположению (2.9) асимптотики растущей размерности при изучении эффекта отбора состоит в том, что r — число отбираемых признаков — пропорционально p, т. е. что

$$r / p \to \rho > 0. \tag{2.52}$$

Естественно также потребовать, чтобы расстояние между классифицируемыми распределениями оставалось ограниченным при росте p и n, т. е. чтобы

$$\sigma^2 = 0\,(p^{-1}). \tag{2.53}$$

Поскольку априори известно, что признаки независимы и нормально распределены с единичной дисперсией, переменную i включаем в число отобранных, когда

$$|\bar{x}_2^{(i)} - \bar{x}_1^{(i)}| > T\sqrt{\sigma^2 + 2/n}, \tag{2.54}$$

где $T = T(\rho)$ определяется из условия $1 - \Phi(T) = \rho/2$. Условие (2.52) выполняется, так как $\bar{x}_2^{(i)} - \bar{x}_1^{(i)} \in N(0, \sigma^2 + 2/n)$. Пусть

$$d^2 = \sum_{i:\,|m_2^{(i)} - m_1^{(i)}| > T\cdot\sigma} (m_2^{(i)} - m_1^{(i)})^2 \equiv \Sigma d_i^2, \tag{2.55}$$

где для i, не удовлетворяющих условию суммирования, положено $d_i = 0$. Согласно (2.3) АОК $P^B_{p,\infty} = \Phi(-\lim d/2)$. Найдем математическое ожидание одного, отличного от нуля, слагаемого в (2.55):

$$\mathbf{E}(d_i^2 \mid d_i^2 > T^2\sigma^2) = \frac{2}{\rho}\int_{-\infty}^{-T\sigma} x^2 \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}dx =$$

$$= \frac{2\sigma^2}{\rho}\left(\Phi(-T) + \frac{T}{\sqrt{2\pi}}\exp\{-T^2/2\}\right). \tag{2.56}$$

Число отличных от нуля слагаемых асимптотически равно ρp, поэтому для больших p

$$d^2 \approx 2\sigma^2 p\left(\Phi(-T) + \frac{T}{\sqrt{2\pi}}\exp\{-T^2/2\}\right). \tag{2.57}$$

Обозначим $\mathbf{E}_i$ $(i=1, 2)$ условное математическое ожидание по наблюдению X при условии, что 1) $X \in F_i$ и 2) обучающая выборка фиксирована Пусть далее в соответствии с предположением (2.53)

$$\sigma^2 \cdot p \to \mu^2 < \infty, \tag{2.58}$$

тогда согласно (2.57) для $\rho = 0{,}125$; 0,25; 0,5 (соответствующие значения T равны 1,53; 1,15; 0,674) для получения $P_\infty^B = 0{,}01$ μ^2 должно быть соответственно равным 43,1; 29,9; 23,3 и для $P_\infty^B = 0{,}10$ $\mu^2 = 13{,}08$; 9,08; 7,08. Для подсчета асимптотического значения УОК по конечной выборке при отборе и обучении надо найти в изучаемой асимптотике (2.9), (2.52), (2.49), (2.50), (2.58) предел отношения

$$\frac{[(\mathbf{E}_1 - \mathbf{E}_2)\, h^B(X)]^2}{\mathbf{E}_1 (h^B(X) - \mathbf{E}_1 h^B(X))^2}. \tag{2.59}$$

Он существует, так как в силу закона больших чисел существуют конечные пределы числителя и знаменателя. Обозначим его $\tilde{d}^2$, тогда

$$\lim \mathbf{E} P_{n,p}^B = \Phi\left(-\tilde{d}/2\right). \tag{2.60}$$

Теперь для значений ρ, λ, P_∞^B, указанных в табл. 2.3, можно найти соответствующие предельные значения $\varkappa^B$ (табл. 2.4).

Таблица 2.4

P_∞	$\lim n/r$	$\rho=0{,}125$	$\rho=0{,}25$	$\rho=0{,}5$
0,10	0,5	3,18	2,78	2,22
	1	2,56	2,19	1,75
	2	1,96	1,69	1,41
	5	1,48	1,30	1,18
	10	1,23	1,18	1,13
0,01	0,5	7,3	4,9	2,9
	1	3,8	2,7	1,9
	2	2,3	1,79	1,42
	5	1,4	1,29	1,16
	10	1,2	1,14	1,07

Качественное соответствие данным табл. 2.3 полное. Однако численно изучаемый эффект более сильно выражен в асимптотической теории.

В близкой постановке ошибку классификации при отборе переменных изучал В. И. Сердобольский [140]. Отличие

от рассмотренной выше задачи состояло в том, что он рассматривал модель блочно-независимых распределений (см. п. 1.1.5 и 2.3.2) с одинаковым числом оцениваемых в блоках параметров (аналог матрицы $\mathbf{I}_p$ в предположении (2.49)) и вместо предположения (2.50) на J_i -расстояния Кульбака между i-ми блоками ($i = 1, \ldots, k$) накладывалось в асимптотике (2.9) условие

$$k^{-1} \sum_{J_i < v/k} 1 \to F(v), \tag{2.61}$$

где $F(v)$ известно. Соответственно отбор переменных проводился по условию $\widehat{J}_i \gtrless c(m)$, где $c(m)$ подбиралось так, чтобы выполнялось условие (2.52).

2.6. Метод структурной минимизации риска

Одна из основных трудностей традиционных алгоритмов дискриминантного анализа состоит в том, что они существенно зависят от субъективных предположений (см. п.2.1.1), которые трудно и не всегда можно проверить по имеющимся данным. Более того, исследователь порой знает, что предположения не верны, а продолжает ими пользоваться, так как они, существенно ограничивая класс возможных решающих правил, успешно работают в данной предметной области. Таким образом, результат применения дискриминантного анализа существенно субъективен, хотя сами алгоритмы полностью формализованы. Поэтому возникло направление исследований, рассчитанное на тех, кто не хотел бы начинать исследование с субъективных предположений. Цель его — не столько предложить эффективные рабочие формулы, сколько развить общую теорию оценивания и на ее основании доказать, что в принципе при $n \to \infty$ можно найти решение и без априорных предположений. В основе нового подхода лежит известное понятие функции потерь. Для простоты изложения ограничимся случаем двух классов, для которого функция потерь

$$Q(\alpha) = \int (y - J(X, \alpha))^2 \, dF(X, y), \tag{2.62}$$

где $\{J(X, \alpha)\}$ — некоторый класс функций, принимающих значения $j = 1, 2$ и зависящих от абстрактного параметра α; $F(X, y)$ — неизвестная функция распределения (X, y). При вычислениях $Q(\alpha)$ заменяют на ее выборочную оценку

$$Q_{\text{эмп}}(\alpha) = \sum_{i=1}^{n} (y_i - J(X_i, \alpha))^2 / n. \tag{2.63}$$

Обозначим класс функций $\{J(X, \alpha)\}$ через S, обычно его выбирают заметно шире, чем класс функций, используемый в статистическом дискриминантном анализе. Это позволяет, с одной стороны, отказаться от априорных предположений, а с другой — служит источником трудностей, о которых речь ниже.

В классе S ищут минимум по α $Q_{\text{эмп}}(\alpha)$. Пусть он достигается при $\alpha = \alpha_0$. Далее строится оценка, показывающая, насколько могут отличаться $Q(\alpha_0)$ и $Q_{\text{эмп}}(\alpha_0)$. Эта оценка зависит от доверительной вероятности ε и имеет вид

$$P\left\{|Q(\alpha_0) - Q_{\text{эмп}}(\alpha_0)| < \Omega\left(\frac{h}{n}, -\frac{\ln \varepsilon}{n}\right)\right\} > 1 - \varepsilon, \qquad (2.64)$$

где $\Omega(u, v)$ — известная непрерывная функция своих аргументов $\Omega(0,0) = 0$; n — объем выборки; h — новое фундаментальное понятие, называемое *емкостью класса* S [44, с. 195—196]. Оно характеризует число способов разделения выборки $X_1, \ldots, X_n$ с помощью функций класса S. Поскольку это понятие не используется в дальнейшем, не будем его здесь определять, а отошлем читателя к оригинальным работам [44, 45]. Отметим только, что емкость $h = p$ в случае линейных от функций X правил вида

$$J(X, \alpha) = U\left(\sum_{i=1}^{p} \alpha_i \varphi_i(X)\right) + 1, \qquad (2.65)$$

где $U(v) = \begin{cases} 1 & v \geqslant 0, \\ 0 & v < 0, \end{cases}$ $\varphi_i(X)$ — известные функции X.

Если бы были сделаны априорные предположения о классифицируемых распределениях, то можно было бы заранее сузить класс функций S, среди которых ищут минимум $Q_{\text{эмп}}$, и формула (2.64) давала бы оценку точности решения. Однако исходная целевая установка заключалась в отказе от априорных предположений и рассмотрении максимально широкого класса S. Для того чтобы соединить потенциальную широту S и ограниченность объема выборки, на S выделяется некоторая структура вложенных друг в друга подмножеств $\{J(X, \alpha)\}$ растущей емкости

$$\begin{aligned} &S_1 \subset \ldots \subset S_q \subset \ldots; \\ &h_1 < \ldots < h_q < \ldots \end{aligned} \qquad (2.66)$$

и минимизация проводится внутри подходящего S_q так, чтобы сбалансировать оцениваемые по обучающей выборке потери от использования не самого широкого класса функций с потерями при переходе от $Q_{\text{окс}}$ к Q, оцениваемыми по

формуле (2.64). Этот подход к построению алгоритмов классификации получил название *структурной минимизации риска*.

Достоинства метода структурной минимизации:

1) отказ от априорных предположений;

2) решение прямой задачи — поиск α_0, а не оценка параметров гипотетических распределений;

3) построение универсальных оценок (2.64);

4) наличие рекомендаций по сочетанию объема выборки n и сложности используемого класса функций;

5) существенное развитие общей теории минимизации эмпирического риска, введение новых понятий, что не может не сказаться на будущем развитии дискриминантного анализа.

Недостатки этого метода:

1) сильно завышены оценки погрешности, делающие метод неконкурентно способным по сравнению с современными алгоритмами дискриминантного анализа;

2) перенос трудностей, связанных с выбором предположений, на этап введения последовательности структур (2.66);

3) отсутствие рекомендаций по выбору структур в зависимости от геометрии расположения классов.

Одна из возможных программных реализаций метода структурной минимизации риска названа алгоритмом «*обобщенный портрет*» [44]. Алгоритм начинается с отображения исходного пространства переменных в бинарное пространство B, каждая координата которого принимает лишь два значения: 0 и 1. Пространство B имеет размерность $p_{\text{нов}} = \sum_{i=1}^{p} k_i$, где k_i — число градаций, на которые разбивается i-й признак. Это обеспечивает универсальность последующей трактовки, а с другой стороны, как показано в п. 2. 3. 4, порой ведет к очень большим потерям информации. Интерпретация формул, получаемых с помощью алгоритма «обобщенный портрет», часто бывает затруднительна из-за большой зашумленности используемых оцифровок.

ВЫВОДЫ

1. В *дискриминантном анализе* (ДА) распределения X в классах известны не полностью. Они задаются *предположениями* и *выборкой*. Обычно предполагается, что либо $f_j(X)$ $(j = 1, \ldots, k)$, либо их отношения принадлежат из-

вестному параметрическому классу функций с неизвестными значениями параметров. Выборка имеет вид $\{(X_i, y_i), i = 1, \ldots, n\}$, где y_i показывает, из какого класса взято наблюдение i.

2. *Алгоритмом ДА* называют метод, с помощью которого на основании обучающей выборки и предположений строится конкретное правило классификации. Поскольку выборка случайна, случайно и построенное на ее основе правило. Поэтому наряду с характеристиками конкретного правила часто рассматривают и средние (ожидаемые) значения этих характеристик, полученные путем усреднения по всем выборкам данного объема n. Это уже характеристика алгоритма. Наиболее часто используются $P^A_{p,n}$ — УОК — *условная ошибка классификации* правила, построенного с помощью алгоритма A при данной обучающей выборке, $\mathbf{E}P^A_{p,n}$ — — ООК — *ожидаемая ошибка классификации* алгоритма A и $P^A_{p,\infty} = \lim\limits_{n\to\infty} \mathbf{E}P^A_{p,n}$ — АОК — *асимптотическая* (при $n \to \infty$) *ошибка классификации* алгоритма A, а также $\varkappa^A = \frac{\mathbf{E}P^A_{p,n}}{P^A_{p,\infty}}$, называемое коэффициентом обучаемости алгоритма A на выборке объема n, или, проще, коэффициентом Раудиса.

3. Для изучения свойств алгоритмов классификации в условиях, когда $p \sim n_j$, удачной оказалась *ассимптотика растущей размерности* Колмогорова — Деева, в которой рассматривается последовательность задач классификации (по параметру m), такая, что $p = p(m)$, $n_j = n_j(m) \to \infty$ и $p/n_j \to \lambda_j < \infty$. Для получения в этой асимптотике содержательных результатов в конкретных задачах на распределения обычно накладываются дополнительные условия.

4. В ДА наиболее часто используются так называемые *подстановочные алгоритмы*, в которых неизвестные в отношении правдоподобия параметры модели заменяются их оценками, построенными по выборке. Пусть α — предельная в асимптотике Колмогорова — Деева минимаксная ошибка классификации. Тогда для подстановочного алгоритма в модели Фишера с известной ковариационной матрицей $\alpha = \Phi(-d^2/2\sqrt{d^2 + \lambda_1 + \lambda_2})$, где d — предельное расстояние между центрами классов: в той же модели, но с неизвестной матрицей $\boldsymbol{\Sigma}$ $\alpha = \Phi(-d^2(1 - \lambda_1\lambda_2/(\lambda_1 + \lambda_2))^{1/2}/2\sqrt{d^2 + \lambda_1 + \lambda_2})$, т. е. заметно больше.

5. Теоретические исследования показывают, что последняя ошибка может быть заметно уменьшена в частных слу-

чаях, когда Σ имеет простую структуру зависимостей Ошибку можно уменьшить также, заменив в линейной дискриминантной функции S^{-1} на специальным образом подобранную регуляризованную оценку $\widehat{S}^{-1}$.

6 В условиях дефицита выборочной информации часто бывает целесообразным для улучшения свойств алгоритма использовать не все переменные, а только часть из них. Вместе с тем задача отбора переменных сопряжена со значительными как техническими, так и чисто статистическими трудностями.

Глава 3. ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО КЛАССИФИКАЦИИ ПРИ НАЛИЧИИ ОБУЧАЮЩИХ ВЫБОРОК (ДИСКРИМИНАНТНЫЙ АНАЛИЗ)

3.1. Предварительный анализ данных

Это один из наиболее ответственных этапов дискриминантного анализа, направленный на формирование математической модели данных, которая в свою очередь служит основой для выбора конкретного алгоритма. Редко исследование с применением ДА осуществляется изолированно. Поэтому при предварительном анализе обязательно надо использовать опыт других близких работ, а не полагаться всецело на данную конкретную обучающую выборку. Кроме того, следует различать условия, при которых метод классификации выводится, и условия, при которых он может быть успешно применен.

Анализ обычно начинается с общего осмотра данных, проводимого с помощью метода главных компонент [11, 10.5]. Ниже описываются более специфические приемы.

3.1.1. Проверка применимости линейной дискриминантной функции (ЛДФ) В п. 1.1 2 ЛДФ выведена как логарифм отношения правдоподобия в задаче Фишера. Соответствующая математическая модель — два многомерных нормальных распределения с общей ковариационной матрицей. Построим графический тест для проверкн этого базового предположения. Но прежде, чем описывать тест, обратим внимание па качественное смысловое различие классов, часто встречающееся в приложениях. Это поможет понять интуитивную идею, лежащую в основе теста. Один из классов обычно соответствует или стабильному состоянию, или устойчивому течению какого-либо процесса. Он относительно однороден. Для него, как правнло, $\pi > 0{,}5$ и нет основания ожидать слишком большого отклонения от многомерной

нормальности распределения X. Назовем объекты этого класса не-случаями С другой стороны, объекты другого класса — случаи — представляют собой отклонения от равновесия, устойчивости Отклонения могут происходить в разных направлениях Можно ожидать, что разброс вектора X для случаев больше, чем для не-случаев Случаи хуже изучены по сравнению с не-случаями

Спроектируем случаи на двумерную плоскость Для этого нормализуем выборочные векторы случаев $X_{1,i}$ согласно выборочным оценкам среднего и ковариационной матрицы не-случаев

$$X_{1,\,i\ \text{норм}} = \mathbf{S}^{-1/2}(X_{1\,i} - \overline{X}_2), \tag{3 1}$$

где X и S определены как обычно

Найдем теперь двумерную плоскость, проходящую через начало координат (центр не-случаев после нормализации), такую, что сумма квадратов расстояний $X_{1,\,i,\,\text{норм}}$ от нее минимальна Нетрудно видеть, что эта плоскость должна быть натянута на первые два собственных вектора, соответствующих наибольшим корням матрицы $\mathbf{B} = \sum_i X_{1,\,i\ \text{норм}} X'_{1,\,i\ \text{норм}}$ Далее спроектируем каждый вектор на эту плоскость и построим отдельно гистограмму, показывающую распределение расстояний случаев от этой плоскости Если n_1 и n_2 достаточно велики по сравнению с p и верны базовые предположения, то линии постоянного уровня плотности случаев должны быть концентрическими окружностями с центром в точке, соответствующей M_1 Распределение расстояний точек $X_{1,i,\,\text{норм}}$ от плоскости должно соответствовать примерно χ^2-распределению с $p - 2$ степенями свободы

Визуальный анализ расположения проекций случаев на плоскости позволяет ответить на следующие вопросы

1 Возможна ли вообще эффективная классификация с помощью плоскости?

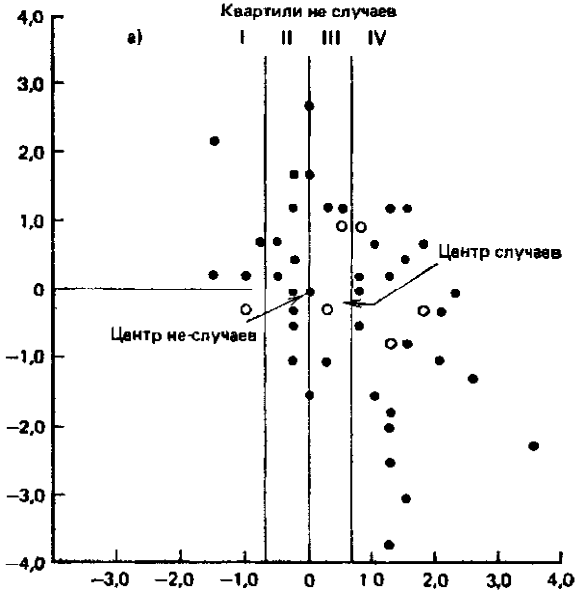
2 Насколько геометрия расположения случаев соответствует гипотезе о равенстве ковариационных матриц?

3 Насколько однородны случаи? Не распадается ли их распределение на отдельные кластеры?

4 Нет ли среди случаев слишком удаленных от плоскости?

и т п

Пример применения предложенного анализа к конкретным данным показан на рис 3 1, *а, б* Из рисунка видно, что: 1) эффективная классификация (в данном случае речь идет о прогнозе события стать случаем) возможна, 2) распределение случаев имеет разброс больше ожидаемого согласно мо-



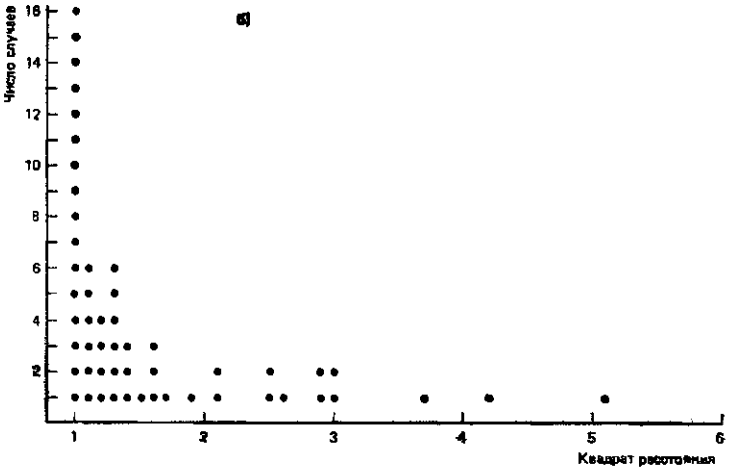


Рис 3 1 Геометрическая проверка условий применимости линейного дискриминантного анализа *а*) проекции случаев на плоскость,

б) распределение квадратов расстояний случаев от плоскости, О — два случая в той же точке

дели двух нормальных распределений с общей ковариационной матрицей; 3) случаи не распадаются на отдельные кластеры. Совпадение распределения расстояний с распределением χ^2_{p-2} вполне удовлетворительно.

3.1.2. «Главные компоненты» одного из классов как новые информативные координаты. Пусть, как и в предыдущем пункте, один из классов из содержательных соображений может быть выделен в качестве стабильного устойчивого состояния и принадлежащие ему объекты названы не-случаями. Объекты других классов будем называть случаями типа j ($j=2, \ldots, k$). Интуитивная идея, лежащая в основе предложения перейти к «главным компонентам» не-случаев, следующая: класс не-случаев не вполне однороден и в него наряду с типичными не-случаями входят объекты, которые все еще остаются не-случаями, но вместе с тем уже сдвинуты в направлениях случаев. Ковариационная матрица не-случаев должна нести следы этих сдвигов. При надлежащей обработке, расположении сдвигов и удаче в выборе параметра λ (см. ниже) эти следы можно выявить и воспользоваться ими при выборе информативных для различения классов координат.

Введем теперь более точные определения. Пусть $\widehat{M}_\lambda$, $\widehat{\mathbf{S}}_\lambda$ — экспоненциально взвешенные оценки среднего и ковариационной матрицы не-случаев [11, п. 10.4.6], причем параметр λ подобран так, чтобы $e(\lambda)$ — средний вес наблюдения обучающей выборки — был бы равен, например, 0,5.
Положим

$$\mathbf{V}_\lambda = \sum_{\text{не-сл}} (X_i - \widehat{M}_\lambda)\, \widehat{\mathbf{S}}_\lambda^{-1}\, (X_i - \widehat{M}_\lambda)' / n_{\text{не-сл}}. \qquad (3.2)$$

Собственные векторы матрицы $\mathbf{V}_\lambda$ будем называть главными компонентами не-случаев. Для нас принципиально важно, что эти компоненты не зависят от векторов других классов. Если при проверке на обучающей выборке окажется, что векторы $\widehat{M}_{\text{сл. типа } i} - \widehat{M}_\lambda$ ($i=2, \ldots, k$) достаточно хорошо описываются первыми главными компонентами, то переход к новым координатам на базе первых главных компонент целесообразен. При проверке удобно построить график «(отношение квадрата проекции вектора на первые l главных компонент к квадрату длины вектора) $\times l$» (рис. 3.2) с нанесенным ожидаемым значением квадрата длины проекции единичного равномерно распределенного случайного вектора, равным l/p, с учетом соответствующих стандартных отклонений, примерно равных $p^{-1}\sqrt{2l\,(1-l/p)}$. Как видно из рисунка, проекции на первые три главные компоненты значимо отличаются от ожидаемого значения.

3.1.3. Устойчивые оценки параметров распределений в классах. Когда распределения X в классах можно считать приближенно многомерными нормальными, для оценки средних и ковариационных матриц рекомендуется использовать устойчивые к отклонениям от нормальности оценки, например ЭВ-оценки [11, п. 10.4.6]. При этом наблюдения, полу-

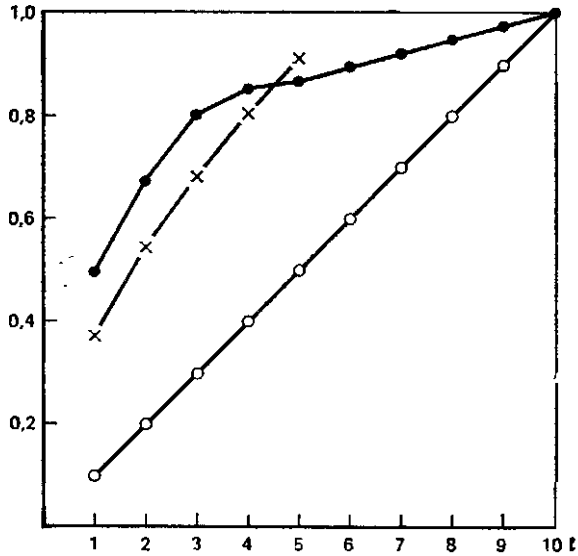


Рис. 3.2. Отклонения квадрата длины проекции вектора на l первых главных компонент от ожидаемого значения при полностью случайной ориентации вектора ($p=10$):
● - отношение квадрата проекции вектора на l первых главных компонент к квадрату длины вектора,
○ — математическое ожидание отношения;
× — математическое ожидание отношения плюс два асимптотических ($p\to\infty$) стандартных отклонения

чившие аномально низкий вес, должны быть внимательно проэкзаменованы: не вкралась ли в их запись (X, y) ошибка. ЭВ-оценки помогают эффективно определять параметры классов при возможных ошибках в отнесении наблюдений к классам.

3.1.4. Проверка гипотез о простой структуре Σ. В § 2.3 показано, что информация о простой структуре ковариационной матрицы Σ дает возможность существенно улучшить результаты классификации. Поэтому всякий раз перед при-

менением ЛДФ следует проверить, имеет ли ковариационная матрица Σ древообразную структуру зависимостей [12, 4.2—4.3]. Для этого с помощью алгоритма Крускала оценивается структура зависимостей, а далее с учетом структуры строится $S_{дсз}$. Пусть S — обычная оценка Σ. Для оценки значимости различия S и $S_{дсз}$ в условиях, когда X внутри класса распределены нормально, можно воспользоваться критерием максимального правдоподобия для проверки сложной гипотезы [11, п. 9.3.3]. При этом если предположение о ДСЗ верно, то величина

$$\gamma = \sum_{i=1}^{n} (X_i - \overline{X})' (S_{дсз}^{-1} - S^{-1}) (X_i - \overline{X}) + n \ln (|S_{дсз}|/|S|), \tag{3.3}$$

где суммирование производится по всем объектам класса, должна иметь асимптотически при $n \to \infty$ χ^2-распределение с $r = p(p+1)/2 - (2p-1)$ степенями свободы, а если гипотеза не верна, то γ должно быть в среднем больше.

3.2. Оценивание отношения правдоподобия

3.2.1. Параметрическое и полупараметрическое оценивание неизвестных плотностей. В том случае, когда на основании априорной информации или предварительного анализа данных можно предположить аналитический вид плотностей распределений в классах, надо использовать обычные подстановочные алгоритмы, следя при этом за тем, чтобы там, где неизвестные параметры в распределениях предполагаются равными, подставлялись одни и те же оценки (см. § 2.3). Ниже разбирается случай полупараметрического оценивания.

Предположим, что имеются два класса с законами распределения $F_j(X) = NT(\Sigma_j, F_j)$, $j = 1, 2$, где $NT(\ldots, \ldots)$ означает класс распределений, трансформируемых к многомерным нормальным (см. п. 1.1.5), для $j = 1, 2$ $F_j(X) = (F_{j1}(x^{(1)}), \ldots, F_{jp}(x^{(p)}))'$ — вектор-функция непрерывных одномерных распределений координат X, Σ_j — положительно определенная матрица. Общая стратегия решения задачи классификации следующая: для каждого класса построить гладкие оценки плотностей $\hat{f}_{ji}$, оценить Σ_j, далее с помощью формулы (1.35) оценить $\hat{f}_j$ и рассмотреть критерий отношения правдоподобия вида $\hat{f}_2/\hat{f}_1 \gtrless c$. Если при постановке задачи сделаны дополнительные предположения, то использовать их при оценке f_{ij} и Σ_j.

Оценивание f_{ji} и F_{ji}. Там, где это ясно из контекста, индексы i и j будем в дальнейшем опускать. Назовем α-квантилем $\left(\frac{1}{n+1} < \alpha < \frac{n}{n+1}\right)$ вариационного ряда $\xi_1 < \xi_2 < \ldots < \xi_n$ эмпирического распределения i-й координаты в выборке из j-й совокупности величины

$$q(\alpha) = (1-\delta)\,\xi_{[\alpha(n+1)]} + \delta\,\xi_{[\alpha(n+1)]+1},$$

где $[u]$ — целая часть u и $\delta = u - [u]$.

Выберем теперь число

$$s = s(n) = \begin{cases} [\log_2 n] + 2 & \text{при } n \leqslant 64 \\ [\sqrt{n}] & \text{при } n > 64 \end{cases}$$

и построим последовательность $(q_0, q_1, \ldots, q_{s+1})$, где для $m = 1, 2, \ldots, s$ $q_m = q(m/(s+1))$, $q_0 = 3q_1 - 2q_2$, $q_{s+1} = 3q_s - 2q_{s-1}$.

Положим теперь

$$\widehat{f}_i(t) = \begin{cases} 0 \text{ для } t < q_0, \\ ((s+1)(q_{m+1} - q_m))^{-1} \text{ для } q_m \leqslant t < q_{m+1}, \\ 0 \text{ для } t > q_{s+1}; \end{cases} \tag{3.4}$$

$$\widehat{F}_i(t) = \begin{cases} 0 \text{ для } t < q_0, \\ \frac{m}{s+1} + (t - q_m)\,\widehat{f}_i(t) \text{ для } q_m \leqslant t < q_{m+1}, \\ 1 \text{ для } t > q_{s+1}. \end{cases} \tag{3.4'}$$

В качестве оценок $\Sigma_j = \|\sigma_{j,ik}\|$ возьмем

$$\sigma_{j,ik} = \sum_{m:y_m = j} \Phi^{-1}(r_{ji}(m)/(n_j+1)) \times$$

$$\times\, \Phi^{-1}(r_{jk}(m)/(n_j+1))/n_j, \tag{3.5}$$

где для $m: y_m = j$ $r_{ji}(m)$ — ранговый номер $x_m^{(i)}$ в вариационном ряду значений i-й координаты в выборке из j-го класса. Заменим в формуле (1.35) неизвестные параметры их оценками и построим оценку отношения правдоподобия.

Если дополнительно предположить, что $\Sigma_1 = \Sigma_2$, то при оценке плотностей надо использовать объединенную оцеику

$$\sigma_{ik} = \frac{n_1}{n_1+n_2}\,\sigma_{1,ik} + \frac{n_2}{n_1+n_2}\,\sigma_{2,ik}. \tag{3.6}$$

Если дополнительно предположить еще, что

$$\Phi^{-1}(F_2(X)) = \Phi^{-1}(F_1(X)) + L, \tag{3.7}$$

где $L=(l^{(1)}, \ldots, l^{(p)})$ — неизвестный вектор (см. п. 1.1.5), то после преобразования координат получаем модель Фишера. В этом случае $T(X)$-объединенную функцию преобразования к нормально распределенным величинам можно найти путем итерационного решения системы уравнений

$$T(X)=\frac{n_1}{n_1+n_2}\Phi^{-1}(\mathbf{F}_1(X))+\frac{n_2}{n_1+n_2}(\Phi^{-1}(\mathbf{F}_2(X))+L); \quad (3.8)$$

$$L=\sum_{m:y_m=2} T(X_m)/n_2-\sum_{m:y_m=1} T(X_m)/n_1. \quad (3.9)$$

3.2.2. Непараметрическое оценивание плотностей. В случае, когда сделать предположение об аналитическом виде $f_j(X)$ нельзя, делают предположение о гладкости $f_j(X)$ и оценивают $\gamma(X)$ как отношение непараметрических оценок плотностей

$$\widehat{\gamma}(X)=\frac{n_2^{-1}\sum\limits_{i:y_i=2} k(\|X_i-X\|^{1/2}/b)}{n_1^{-1}\sum\limits_{i:y_i=1} k(\|X_i-X\|^{1/2}/b)}, \quad (3.10)$$

где $\|Z\|$ — норма элемента Z; b — малый параметр; $k(u)$— функция, удовлетворяющая следующим условиям: $k(u)\geqslant 0$, $\int k(u)\,du=1$, $k(u)\to 0$ ($|u|\to\infty$), $k(u)=k(-u)$. В качестве $k(u)$ обычно берут плотность нормального закона с параметрами (0,1). Наряду с формулой (3.10) широко используется оценка

$$\widehat{\gamma}(X)=\frac{n_2^{-1}\sum\limits_{i:y_i=2}\prod\limits_{j=1}^{p} k\left(\left(x_i^{(j)}-x^{(j)}\right)/b^{(j)}\right)}{n_1^{-1}\sum\limits_{i:y_i=1}\prod\limits_{j=1}^{p} k\left(\left(x_i^{(j)}-x^{(j)}\right)/b^{(j)}\right)}, \quad (3.11)$$

получившая название *оценки Парзена*. Часто для упрощения проводится предварительная покоординатная нормализация переменных, чтобы они имели одну и ту же меру разброса, и b выбираются равными [132].

Для оценок (3.10) и (3.11) ключевым является выбор параметров $b^{(j)}$. Его естественно связать с какой-либо мерой качества классификации (см. п. 1.3.4) аналогично тому, как для задачи регрессии это сделано в [12, § 10.1]. На практике оценки Парзена работают хорошо. Их существенные недостатки: необходимость запоминания всей обучающей последовательности и высокая чувствительность метода к непредставительности обучающей выборки.

В [198] для распределений, несколько похожих на многомерные нормальные, рекомендуется следующая эвристическая приближенная процедура, основанная на рангах. Для каждой из координат $l = 1, \ldots, p$ строится вариационный ряд из n значений $x_l^{(l)}$ [11, п. 5.6.4]. Исходная величина $x_l^{(l)}$ заменяется на $z_l^{(l)}$ — ее номер в вариационном ряду. Если в вариационном ряду были связи, т. е. $\ldots < x_{i_1}^{(l)} = x_{i_2}^{(l)} = \ldots = x_{i_m}^{(l)} < \ldots$, то $z_{i_1}^{(l)} = \ldots = z_{i_m}^{(l)}$ и равняется среднему рангу $x_{i_1}^{(l)}, \ldots, x_{i_m}^{(l)}$ в вариационном ряду. Далее $\{Z_k : y_k = 1\}$ и $\{Z_k : y_k = 2\}$ рассматриваются как выборки из многомерных нормальных совокупностей и классификация проводится по одному из правил для многомерных нормальных распределений. Сравнения этой рекомендации с изложенным в предыдущем пункте подходом с T-нормальными распределениями не проводилось. Однако последний нам кажется более логичным.

3.2.3. Прямое оценивание отношения правдоподобия. Часто аналитический вид плотностей f_j неизвестен, но известен с точностью до неизвестных параметров аналитический вид отношения правдоподобия. Так, в частности, будет, если в модели Фишера каждое из наблюдений обучающей выборки удаляется или остается в выборке независимо от других наблюдений с вероятностью, зависящей только от значения X. В этом частном случае

$$f_j(X) = g(X) \cdot \varphi(X, M_j, \Sigma) \quad (j = 1, 2), \tag{3.12}$$

где φ — плотность многомерного нормального закона; $g(X)$ — некоторая неизвестная положительная функция, вообще говоря, зависящая от π_j, M_j, Σ. Несмотря на то что (3.12) может заметно отличаться от плотности нормального закона, отношение правдоподобия по-прежнему остается линейной функцией X:

$$h(X) = \ln \gamma(X) = (X - (M_1 + M_2)/2)' \Sigma^{-1} (M_2 - M_1) =$$

$$= X' \Theta + \theta. \tag{3.13}$$

Условная вероятность гипотезы H_1, когда дано наблюдение X, легко выражается через $h(X)$:

$$P\{H_1 | X\} = \pi_1 f_1(X) \Big/ \left(\sum_j \pi_j f_j(X)\right) = \left(1 + \frac{\pi_2}{\pi_1} \gamma(X)\right)^{-1} =$$

$$= (1 + \exp\{\ln(\pi_2/\pi_1) + h(X)\})^{-1} \equiv p(X). \tag{3.14}$$

В частном случае, когда $h(X)$ — линейная, как в (3.13), функцня от X

$$p(X) = (1 + \exp\{\theta^{(0)} + X'\Theta\})^{-1}, \qquad (3.15)$$

где $\theta^{(0)} = \theta + \ln(\pi_2/\pi_1)$. Функция, стоящая в правой частн (3.15), называется *логистической*.

Предполагая, что имеет место (3.15), можно воспользоваться соотношением (3.14) для того, чтобы найти неизвестные параметры $\theta^{(0)}$ и Θ. Для этого воспользуемся методом условного максимального правдоподобия:

$$(\widehat{\Theta}, \widehat{\theta^{(0)}}) = \arg\sup_{\Theta, \theta^{(0)}} \ln L(\{y_k\} | \Theta, \theta^{(0)}, \{X_k\}), \qquad (3.16)$$

где $L = \prod_k p^{2-y_k}(X_k, \Theta, \theta^{(0)})(1-p(X_k, \Theta, \theta^{(0)})^{y_k-1}$.

При условии, что имеет место модель Фишера, метод условного максимального правдоподобия использует не всю информацию, содержащуюся в обучающей выборке. Однако, как показывает теоретическое исследование [220], проигрыш в эффективности для близких совокупностей незначителен.

В случае, если на обучающей выборке совокупности могут быть отделены друг от друга некоторой плоскостью, максимальное значение $\ln L$ равно бесконечности и решенне уравнения (3.16) не единственно. Тогда надо просто найти соответствующую плоскость, например с помощью метода потенциальных функций (см. п. 1.3.3). Рекомендации, как действовать в случаях, когда при некоторых значениях аргумента $\ln L \to \infty$, можно найти в [175].

3.2.4. Непараметрическое оценивание отношения правдоподобия. Наиболее известен здесь метод «k-ближайших соседей», предложенный в работе [225]. Он состоит в следующем:

1) в пространстве наблюдений вводится расстояние между произвольными точками $\rho(X_1, X_2)$;

2) в зависимости от объема обучающей выборки n и предположений о гладкости плотностей распределения классифицируемых совокупностей выбирается нечетное k;

3) вокруг классифицируемой точки Z строится сфера $O_k(Z)$ наименьшего радиуса ρ, содержащая не менее k точек из обучающей последовательности;

4) точка Z относится к той совокупности, к которой принадлежит большинство точек из обучающей выборки, попавших в $O_k(Z)$.

Конечно, вместо сфер можно было бы брать области более общего вида. Например, фиксировать какую-либо

окрестность нуля U ограниченного диаметра и рассматривать системы окрестностей вида $U_\rho(Z) = \{X : Z - X = rV$, где $V \in U$ и $r \leqslant \rho\}$, ρ — произвольное положительное число.

Некоторые теоретические вопросы, связанные с изложенным методом, обсуждаются в [108].

3.2.5. Локальная линейная аппроксимация отношения правдоподобия. В [12, п. 10.1.4 и § 10.2] видим, что в регрессионных задачах эффективным оказывается использование локальных параметрических описаний регрессии. По сравнению с традиционным непараметрическим подходом оно в меньшей степени зависит от особенностей обучающих выборок и позволяет получить более полное описание регрессионной поверхности. Аналогично и в задаче классификации. Пусть X_0 — произвольная точка, тогда правдоподобно, что в достаточно широкой ее окрестности $O(X_0)$ приближенно выполняется соотношение

$$h_{X_0}(X) \approx \theta + \Theta'(X - X_0). \tag{3.17}$$

Оценка параметров этой модели на $X_i \in O(X_0)$ позволяет не только провести классификацию нового наблюдения в точке $X = X_0$ по значениям $\widehat{\Theta}$, $\widehat{\theta}$ и отношению $\widehat{\pi}_2/\widehat{\pi}_1$, где $\widehat{\pi}_j$ — доля наблюдений в обучающей выборке из j-й совокупности в окрестности $O(X_0)$, но и получить описание отношения правдоподобия в окрестности X_0.

3.3. Сводка рекомендаций по линейному дискриминантному анализу

Линейным дискриминантным анализом (ЛДА) называют совокупность алгоритмов, связанных с общей моделью Фишера (см. п. 1.1.2) и некоторыми ее обобщениями, сохраняющими общий линейный (по X) вид решающего правила (1.12), (2.15), п. 2.4.3, (3.15).

3.3.1. Проверка базовых предположений. Заметим сначала, что выраженная негауссовость одномерных распределений $x^{(k)}$ $(k = 1, \ldots, p)$ при гипотезах H_j $(j = 1, 2)$ (например, дискретность распределений) обычно не рассматривается в качестве серьезной помехи к применению линейной дискриминантной функции (ЛДФ). Более важны другие свойства модели: существование постоянных a_i, таких, чтобы $f_1(x^{(i)} + a_i) \approx f_2(x^{(i)})$ и f_j были бы примерно симметричны. Или даже просто выполнение условий $\mathbf{E}(x^{(i)} | H_1) \neq \mathbf{E}(x^{(i)} | H_2)$ и $\mathbf{D}(x^{(i)} | H_1) \approx \mathbf{D}(x^{(i)} | H_2)$.

Описанный в п. 3.1.1 визуально-графический метод дает комплексную проверку условий применимости ЛДА. Если распределения ни одного из классов не распадаются на отдельные кластеры, то можно попытаться добиться большего совпадения с моделью Фишера с помощью параметрического преобразования координат [11, п. 10.3.4] или перехода к T-нормальным распределениям (см. пп. 1.1.5 и 3.2.1).

3.3.2. Гипотеза о простой структуре зависимостей между признаками. Примеры распределений с простой структурой связей даны в пп. 1.1.2 и 1.1.5. Независимость признаков, наличие ДСЗ или $R(k)$ позволяют путем использования оценок $\mathbf{S}^{-1}$, учитывающих структуру зависимостей, заметно уменьшить ООК (см. § 2.3). Кроме того, в этом случае отбор информативных признаков носит неитеративный характер и всегда можно сказать, почему включен или не включен в число отобранных тот или иной признак (см. п. 1.4.1). Метод проверки гипотезы о наличии ДСЗ описан в п. 3.1.4. Эту проверку целесообразно проводить всегда.

3.3.3. Методы выделения информативных комбинаций координат. Линейные комбинации — это главные компоненты общей ковариационной матрицы данных или главные компоненты, связанные с ковариационной матрицей одного из классов (см. п. 3.1.2). Последние легче интерпретировать, так как в них направления компонент статистически не зависят от средних второй совокупности. Иногда бывает целесообразным выделить, исходя из содержательных соображений, подгруппу $X^{(l)}$ координат, направленных на оценку только одного прямо не измеримого свойства объекта, и спроектировать $X^{(l)}$ на направление первой главной компоненты этой подгруппы для наибольшего класса. Обозначим проекцию $z^{(l)}$. Замена $X^{(l)}$ на $z^{(l)}$ позволяет существенно сократить размерность пространства переменных, учитываемых в асимптотических формулах § 2.3.

3.3.4. Методы вычислений. Если $\min\limits_{j=1,2} n_j \gg p$, то в случае, когда распределения X близки к многомерным нормальным законам, можно использовать подстановочный алгоритм (см. п. 2.1.1), в остальных случаях лучше подгонять логистическую функцию (см. п. 3.2.3), как менее зависящую от гауссовости распределений. В случае $\min\limits_{j=1,2} n_j \sim p$, когда нельзя сделать упрощающих предположений о зависимости координат (см. 3.3.2), целесообразно для уменьшения ООК использовать регуляризованные оценки $\mathbf{S}^{-1}$ (см. § 2.4).

3.3.5. Альтернативные алгоритмы. Если исходные предположения ЛДА не выполняются (см. п. 3.3.1) и их выполне-

ния нельзя добиться преобразованием координат или переходом к T-нормальному варианту модели Фишера (п.3.2.1), можно рекомендовать либо малопараметрические представления распределений в виде смесей (см. п. 1.1.3), либо использовать непараметрические методы пп.3.2.2 и 3.2.4.

3.3.6. Другие вопросы. В случае, когда предположения о простой структуре зависимостей не верны, *отбор информативных переменных* проводится с помощью общего подхода, изложенного в п. 1.4.3. Полученный результат контролируется при этом обычно с помощью оценки расстояния Махалаиобиса (см. п. 3.4.3) и с учетом эффектов, описаниых в § 2.5.

В случае, когда есть подозрение, что некоторые наблюдения в обучающей выборке (X_i, y_i) могут быть определены с ошибкой (*засоренные выборки*), надо использовать устойчивые оценки параметров распределений, как это рекомендуется в п. 3.1.3.

3.4. Оценка качества дискриминации

Как сказано в § 2.1, оценка качества построенного правила классификации является завершающей операцией ДА. Выбор конкретных показателей и методов их оценивания зависит от целей построения правила классификации, от начальных предположений и степени уверенности в них, от выбранного алгоритма и, наконец, от доступного программного обеспечения.

3.4.1. Показатели качества разделения. В табл. 3.1 дана сводка основных показателей качества дискриминации, там же указано, где в книге можно найти соответствующие разделы. Средняя ошибка входит в две группы показателей (1.2 и 2.1). Показатели (1.3 и 3.1) так же связаны друг с другом. Их сопоставление может быть использовано для прямой проверки применимости модели Фишера. Особое место занимают показатели, требующие численной оценки отношения правдоподобия в каждой точке выборочного пространства (2.2 и 3.2). Если умеем его оценивать, то «первичиая» оценка расстояния Бхатачария по обучающей выборке может выглядеть, иапример, следующим образом:

$$B = \int (f_1 f_2)^{1/2}\, dX \approx \frac{1}{n_1+n_2}\left(\sum_{i:y_i=1}\left(\frac{\widehat{f_2(X_i)}}{f_1(X_i)}\right)^{1/2} + \right.$$

$$\left. + \sum_{i:y_i=2}\left(\frac{\widehat{f_1(X_i)}}{f_2(X_i)}\right)^{1/2}\right).$$

Таблица 3.1

№ п/п	Класс показателей	Показатели	Где описаны
1	Ошибка классификации	1.1. Минимаксная ошибка 1.2. Средняя ошибка 1.3. Кривая «чувствительность — специфичность» и расчетные значения d	п. 1.2.4, ф-ла (1.40) п. 1.1.4, ф-ла (1.31) п. 1.2.2
2	Функция потерь	2.1. Средняя ошибка 2.2 То же, но с взвешенным учетом промежуточных значений отношения правдоподобия	п. 1.1.4, ф-ла (1.31′) п. 1.1.4
3	Расстояние между распределениями	3.1. Расстояние Махаланобиса d^2 3.2. Расстояния, определяемые через отношение правдоподобия	п. 1.2.4, ф-ла (1.39) п. 1.2.4, ф-лы (1.42) и (1.43)

Смысл слова «первичная» будет ясен из материала следующего пункта.

3.4.2. Методы оценивания. Хорошо известно, что если применить построенное правило классификации к обучающей выборке, то оценка качества классификации будет в среднем завышена по сравнению с той же оценкой качества по не зависимым от обучения данным. Это означает, что регистрируемые на обучающей выборке значения ошибок и функции потерь будут ниже ожидаемых, а значения расстояний — больше. Укажем основные приемы борьбы с этим завышением качества.

Разбиение имеющихся данных на две части: *обучающую* и *экзаменующую выборки*. Это самый простой и убедительный метод. Им следует широко пользоваться, если данных достаточно. Тем более что, если разбиение данных произведено по какому-либо моменту времени, метод позволяет оценивать качество правила, построенного по прошлым данным, в применении к сегодняшним данным. С чисто статистической точки зрения метод разбиения данных на две части расточителен. Поэтому предложен ряд других, более сложных методов, которые полнее используют выборочную информацию.

Метод скользящего экзамена. При этом методе одно из наблюдений отделяется от выборки и рассматривается в ка-

честве экзаменующего наблюдения. По оставшимся $n-1$ наблюдениям строится правило классификации, которое применяется к выделенному наблюдению. Результат применения регистрируется и оценивается. Наблюдение возвращается в выборку, выделяется следующее наблюдение и т. д. Процесс прекращается через n шагов, когда будет перебрана вся выборка. Последовательные оценки, получаемые с помощью скользящего экзамена, несмещены, однако зависимы между собой. Существенная особенность метода — n-кратное построение правила классификации. В случае непараметрических оценок пп. 3.2.2 и 3.2.4 это сделать легко — достаточно просто не включать выделенное наблюдение в суммы в формулах (3.10), (3.11) или не учитывать его в окрестности $O(X)$. В случае использования линейной дискриминантной функции, оцениваемой через $\overline{X}_1$, $\overline{X}_2$, $\mathbf{S}^{-1}$, при коррекции $\mathbf{S}^{-1}$ используется формула Бартлетта для обратных симметричных матриц $\mathbf{A}$

$$(\mathbf{A}+UV')^{-1}=\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}UV'\mathbf{A}^{-1}}{1+U'\mathbf{A}^{-1}V}, \tag{3.18}$$

которая существенно упрощает расчеты. В общем случае, особенно при отборе переменных, метод скользящего экзамена слишком трудоемок.

Использование обучающей выборки в качестве экзаменационной с последующей поправкой на смещение. Идея метода достаточно проста. Пусть оценивается некоторый параметр r. Обозначим его оценку на обучающей выборке $\widehat{r}_{\text{об}}$ и оценку на новой выборке $\widehat{r}_{\text{экз}}$. Пусть далее $\Delta = \mathbf{E}(\widehat{r}_{\text{экз}}-\widehat{r}_{\text{об}})$, а $\widehat{\Delta}$ — некоторая оценка Δ. Тогда

$$\widehat{r}=\widehat{r}_{\text{об}}+\widehat{\Delta}. \tag{3.19}$$

Предложены различные способы оценки Δ: аналитические, опирающиеся на предельные соотношения гл. 2, и эмпирические, использующие специальные вычислительные процедуры. Оба подхода описываются ниже.

3.4.3. Аналитические поправки. Они наиболее просты в вычислительном плане, но существенно опираются на математические предположения проверяемых моделей. Поэтому их следует рассматривать только в качестве первых приближений.

Поправка для оценки расстояния Махаланобиса в модели Фишера. Пусть

$$D^2=(\overline{X}_1-\overline{X}_2)'\,\mathbf{S}^{-1}(\overline{X}_1-\overline{X}_2). \tag{3.20}$$

Оценка D^2 смещена. Несмещенная оценка расстояния Махаланобиса [264]

$$\widehat{d^2} = \frac{n-p-3}{n-2} D^2 - \left(\frac{p}{n_1} + \frac{p}{n_2}\right). \tag{3.21}$$

Поправка для ООК. На основании теоретического рассмотрения модели Фишера и ряда результатов моделирования с различными алгоритмами Раудис Ш. [132] рекомендует при конструировании поправки использовать параметр $\varkappa$ (см. гл. 2); если $\varkappa < 1{,}5$ и $\widehat{\alpha}_R$ — оценка ошибки классификации, полученная на обучающей выборке, то $\widehat{\alpha}$ — оценка ООК может быть приближенно оценена с помощью

$$\widehat{\alpha} = \varkappa^2 \cdot \widehat{\alpha}_R. \tag{3.22}$$

3.4.4. Метод статистического моделирования (bootstrap method). Предложен В. Эфроном [219]. В нем рекомендуется принять обучающую выборку за генеральную совокупность. Из нее производить повторные по параметру i наборы обучающих и экзаменующих выборок и для каждой i-й пары выборок оценивать разность $\Delta_i = r_{i,\text{экз}} - r_{i,\text{об}}$. Среднее арифметическое Δ_i принимается за $\widehat{\Delta}$. Далее используется формула (3.19).

3.5. Рекомендации для $k \geqslant 3$ классов

Условимся говорить, что объем (обучающей) выборки достаточен, если $\min\limits_{j=1,\ldots,k} n_j \gg p$; объем выборки относительно ограничен, если $\min\limits_{j=1,\ldots,k} n_j \sim p$. Знаки $\gg$ (много больше) и $\sim$ (одного порядка) здесь надо трактовать с учетом геометрии расположения классов: при пересечении центральных частей распределений X в разных классах наблюдений должно быть больше, а при упорядоченности центральных частей распределений вдоль какой-либо гладкой кривой — меньше. В первом случае выборочные оценки параметров распределений при аналитических предположениях о форме распределений (см. п. 3.2.1) и прямые описания распределений при использовании полупараметрических и непараметрических методов (см. пп. 1.3.2 и 3.2.2) дают довольно хорошие совпадения реальных распределений и их оценок, поэтому в этом случае полностью можно использовать материал п. 1.5.

Во втором случае, когда объем выборки относительно ограничен, надо использовать упрощающие предположения, но только такие, которые не выводят за рамки распределений. В частности, следует опасаться использовать поправочные члены, возникающие из предельных теорем в традиционной асимптотике теории вероятностей, хотя такие предложения порой и вносятся [184]. Описываемые ниже предположения упорядочены по степени ограничений, накладываемых на распределения в классах: сначала идут наиболее сильные предположения, затем они ослабляются. Конечно, полной упорядоченности достичь не удается, так как ограничения существенно не одномерны.

Независимость одномерных распределений координат в классах. Это предположение довольно часто и успешно использовалось при диагностике в случае большого числа классов.

Но сегодня его следует заменить на более реалистическое предположение, что переменные в классах имеют древообразную структуру зависимостей (см. пп. 1.2.2, 1.1.5, 2.3.3). При этом в случае предположений о нормальных распределениях в классах можно требовать совпадения соответствующих ковариационных или (что не то же самое) корреляционных матриц или вообще ограничиться требованием, чтобы одинаковым в классах был только граф структуры зависимостей [12, § 4.2—4.3], а ковариационные матрицы различны. Наконец, можно потребовать равенства ковариационных матриц, не предполагая ДСЗ.

Предположения о средних: 1) средние лежат в пространстве первых главных компонент одного из классов (см. п. 3.1.2); 2) средние классов лежат на прямой; 3) классы могут быть упорядочены (см. п. 1.5.3).

ВЫВОДЫ

1. Предварительный анализ данных — один из наиболее ответственных этапов дискриминантного анализа. При его проведении следует различать условия, при которых конкретное правило классификации выводится, и условия, при которых оно применяется. Так, теоретическим основанием для *линейной дискриминантной функции* служит модель Фишера, применяется же ЛДФ в значительно более широких условиях.

2. Основные методы ДА основаны на параметрических, полупараметрических и непараметрических оценках плотно-

стей распределенний или на непосредственной оценке отношения правдоподобия.

3 В настоящее время еще не решена задача создания единого дерева рекомендаций по проведению ДА, полностью исключающего субъективный фактор. Поэтому рекомендации приходится группировать по разделам: проверка базовых предположений, упрощающих условий, методы вычислений, альтернативные решения и т.п. с неформалнзованным выбором между альтернатнвамн.

4. Оценки качества конкретного правила классификации проводятся либо на новой выборке, либо на обучающей выборке Первый метод дорог, но наиболее убедителен. Во втором случае, чтобы избежать искусственного улучшения результатов, либо к параметру качества, оценеиному путем рекласснфикации обучающей выборки, примеияется поправка, полученная аналитическн или с помощью метода математического моделирования, либо используется *метод скользящего экзамена*. Последиий состоит в том, что одио из наблюдений исключается из выборки, по оставшимся строится правило классификации, которое применяется к первому наблюдеиию, затем первое наблюдение возвращается в выборку и исключается второе, по оставшимся наблюдениям строится новое правило классификации и примеияется ко второму выделенному и так далее до тех пор, пока не будут по очереди классифицированы все наблюдеиия. По итогам классификации строится оценка качества.

Глава 4. ПРИМЕНЕНИЯ ДИСКРИМИНАНТНОГО АНАЛИЗА

Статистические методы классификации применяются при распознавании сигналов, диагностике состояний сложных технических систем и человека, а также при прогнозировании будущих отказов, иеисправностей, заболеваний. Использоваиие статистических методов для решения принципиальио новых для конкретной области зиаиия задач всегда носит творческий характер и часто требует приспособления и развития соответствующего математического аппарата. Поэтому при изложении материала большое внимание уделяется как методическим особенностям применения описанных в предшествующих главах методов, так и изложению математического инструментария, иаправленного на решение тех же задач, что и классификация, с обязательным указанием связи между методами.

4.1. Группы риска и сравнительные испытания

4.1.1. Группы риска. Пусть группа объектов периодически подвергается осмотру с целью обнаружения неисправных объектов, а также выделения объектов, которые исправны в момент осмотра, но могут выйти из строя до следующего осмотра. Для решения поставленной задачи, если, конечно, нет прямых надежных индикаторов возникновения в будущем неисправности, можно воспользоваться методом статистической классификации. Пусть X — результат осмотра исправного объекта. Тогда на основании значения X можно попытаться принять одно из двух решений (гипотез): H_1 — «объект останется исправным до следующего осмотра» или H_2 — «объект выйдет из строя до следующего осмотра». Если условные распределения $P(X \mid H_1)$ и $P(X \mid H_2)$ основательно пересекаются, а это типичный случай, то ошибки классификации (см. § 1.2) будут высокими и такой подход индивидуального предсказания судьбы объекта малопродуктивен. Вместе с тем можно оценить $P\{H_2 \mid X\}$ и тем самым отнести соответствующий объект к одной из групп риска H_2. Такой прогноз, в отличие от первого, иногда называют групповым (не путать с групповой классификацией). Оба метода прогноза почти не отличаются по используемому математическому аппарату, различны лишь формы представления результатов (см. § 1.2). Однако с точки зрения приложения они принципиально различны. Нечетким предсказанием индивидуальной судьбы объекта (в терминах H_1 и H_2) воспользоваться трудно. В то же время указание группы риска весьма информативно. В самом деле, если есть ограниченный дополнительный ресурс для более полного обследования объектов, то его, видимо, целесообразно применить к объектам, принадлежащим к группам более высокого риска. Так, например, поступают при диспансеризации населения. При лечении профилактические средства с заметным побочным действием также стоит давать только тем больным, у которых ожидаемый основной эффект лекарства будет выше ожидаемого ущерба от побочных действий, т. е. и здесь учет $P\{H_2 \mid X\}$ крайне существен.

В разобранной выше задаче лишь немного отклонились от традиционной формы представления результатов и сразу же получили очень интересные варианты практического использования ДА.

4.1.2. Индикаторы и факторы риска. Предположим, что в разобранной в предыдущем пункте задаче хотим найти компоненты X, наиболее тесно связанные с осуществлением события H_2. С помощью описанных в предыдущих главах

методов (см. § 1.4, 2.5) можем выделить группу переменных $\tilde{X} = (x^{(l_1)}, \ldots, x^{(l_k)})'$, такую, что сила прогноза при расширении набора $\tilde{X}$ до исходного X на имеющемся в распоряжении материале статистически значимо не увеличивается. Переменные, входящие в $\tilde{X}$, называют *риск-индикаторами* H_2. При этом в слове индикатор выделяются два смысловых оттенка: 1) на индикатор не всегда можно воздействовать, например, как на возраст объекта и 2) индикатор не обязательно причинно обусловливает возникновение H_2. Он, например, может быть только связан с внутренним механизмом, порождающим H_2.

Перевод части индикаторов в факторы риска. Предположим, что можно воздействовать на часть риск-индикаторов, например $x^{(l_1)}, \ldots, x^{(l_m)}$ $(m < k)$, изменяя их значение на новые $x_н^{(l_1)}, \ldots, x_н^{(l_m)}$, в то время как остальные риск-индикаторы остаются без изменения. Обозначим $X_{i,н}$ вектор риск-индикаторов для i-го объекта после изменения. Если после различных воздействий частота события H_2 останется сопоставимой с $\sum_i P\{H_2 \mid X_{i,н}\}/\sum_i 1$, где условная вероятность подсчитывается по установленным ранее для X формулам и профессиональный анализ показывает, что переменные $x^{(l_1)}, \ldots, x^{(l_m)}$ можно рассматривать как непосредственные составляющие механизма возникновения H_2, то эти переменные называют *риск-факторами* H_2. На этом пути были, в частности, установлены риск-факторы развития ишемической болезни сердца, послужившие основой развертывания широкой программы профилактики сердечно-сосудистых заболеваний [277, 322].

4.1.3. Сравнительные испытания. Предположим, что к описанным в п.4.1.1 объектам, признанным исправными при осмотре, применяются определенные воздействия с целью предотвратить их выход из строя за определенный промежуток времени. Для того чтобы эмпирически отобрать наиболее эффективное воздействие, проводятся так называемые *сравнительные испытания*. В простейшем случае они заключаются в следующем. Пусть требуется сравнить два воздействия: A — старое и B — новое. Из объектов образуются две по возможности близкие по свойствам $\{X_i\}$ группы: O — основная и K — контрольная. К объектам основной группы применяется воздействие B, а к объектам контрольной группы — воздействие A. Об эффективности воздействий судят по альтернативному признаку: остался ли объект исправным (событие H_1) или вышел из строя (собы-

тие H_2). Вопросам формирования сравниваемых групп посвящена обширная статистическая литература [85, 102]. Тем не менее добиться полного сходства групп даже при умеренной размерности X удается редко. Это обстоятельство мешает интерпретации результатов испытаний, поскольку априори известно, что $P\{H_2|X\}$ зависит от X.

В случае, когда заранее известны риск-группы при старом воздействии A, поправку на неоднородность основной и контрольной групп сделать не трудно. Для этого достаточно оценить разность

$$\delta = P\{H_2 | A,\ X \in O\} - P\{H_2 | A,\ X \in K\} \qquad (4.1)$$

и далее проверять гипотезу, что

$$H_0\ \ P\{H_2 | B,\ X \in O\} = P\{H_2 | A,\ X \in K\} + \delta. \qquad (4.2)$$

Частным, но практически важным случаем «испытаний» является анализ эффективности разных воздействий на ретроспективных данных. Возможность такого анализа обусловлена тем, что четкие однозначные правила назначения воздействия в зависимости от X обычно или отсутствуют, или в силу разных причин не соблюдаются и поэтому в банках данных накапливается довольно обширная информация о различных сочетаниях пар (X, воздействие) и соответствующих исходах. Многочисленные примеры проведенных исследований показывают, что на основании априорных профессиональных соображений исследователь может разделить объекты на относительно однородные группы риска — страты и проводить анализ эффективности внутри соответствующих групп [85, 179]. Видимо, целесообразно включать проведение подобного анализа в качестве специальной задачи информационных технологических систем с целью автоматизированного подбора гипотез для дальнейшего их анализа исследователем.

В случае, когда риск-группы априори не известны и не могут быть убедительно назначены исследователем, приходится рассматривать полную математическую модель ситуации.

Простейшая модель влияния X и воздействия $V \in \{A, B\}$ на условную вероятность H_2 имеет вид:

$$P\{H_2 | X,\ V\} = (1 + \exp\{\theta^{(0)} + \Theta' X + q(V)\})^{-1}, \qquad (4.3)$$

где $q(A) = -q(B) = q$, $\theta^{(0)}$, Θ — неизвестные параметры. Проверяемая в испытании гипотеза заключается в том, что эффект сравниваемых воздействий тождествен, т. е. что

$$H_0 : q = 0. \qquad (4.4)$$

Очевидно, при $q < 0$ более эффективно новое воздействие, а при $q > 0$ — старое. Предположения (4.3) и (4.4) надо дополнить предположениями, что при заданных X и V результаты испытаний независимы и что распределения X в основной и контрольной группах независимы между собой, и задать эти распределения. Например, положив, что в основной группе

$$X \in N(M_O,\ \Sigma), \tag{4.5}$$

а в контрольной

$$X \in N(M_K,\ \Sigma), \tag{4.6}$$

где M_O, M_K, Σ — неизвестные параметры, причем $\det \Sigma \neq 0$. Базовые предположения (4.3) — (4.6) погрузим в одну из асимптотик: традиционную или растущей размерности (см. п.2.2.1). Можно также пополнить модель упрощающими предположениями о взаимной близости векторов M_O и M_K и о структуре Σ.

Сводку практических рекомендаций по методам интерпретации результатов сравнительных испытаний с учетом возможного несовпадения распределений в контрольной и основной группах можно найти в [179].

4.2. Методы описания риска развития события

4.2.1. Мгновенный риск и факторизация Кокса. В предыдущем параграфе для описания вероятности возникновения неисправности за время от одного осмотра до другого использовалось понятие риск-группы. Но для той же цели можно использовать понятие мгновенного риска (или просто риска)

$$r(t) = \lim_{\Delta \to 0} \Delta^{-1} P\{\text{объект неисправен в } t + \Delta \mid \text{объект исправен в } t\}. \tag{4.7}$$

Риск и вероятность события $H = \{$появление неисправности за интервал $s < t < s + T\}$ связаны соотношением

$$P\{H \mid \text{объект исправен до } s\} = 1 - \exp\left\{-\int_s^{s+T} r(t)\,\mathrm{d}t\right\}. \tag{4.8}$$

По аналогии с (4.7) можно ввести условный риск в момент t при условии, что в момент осмотра $s < t$ объект имел вектор показателей $X(s)$ $r(t \mid X(s)) = \lim_{\Delta \to 0} \Delta^{-1} P\{$неисправен в $t + \Delta \mid$ исправен в t, $X(s)\}$.

Понятие условного риска — более тонкий инструмент для описания закономерностей возникновения неисправности, чем $P\{H_2 \mid X(s)\}$ — понятие условной вероятности. Однако $r(t \mid X(s))$, вообще говоря, требует для своей оценки заметно большего числа наблюдений.

С целью частичного преодоления этой трудности в 1972 г. Д. Кокс [206] предложил факторизовать $r(t \mid X(s))$ путем представления

$$r(t \mid X(s)) = g(X(s)) \cdot h(t) \tag{4.9}$$

или

$$r(t \mid X(s)) = g(X(s)) \cdot h(t-s), \tag{4.9'}$$

где $h(t)$ в (4.9) — функция «возраста» объекта, а в (4.9')— функция времени, прошедшего после осмотра; $g(\ldots)$ — функция изучаемых признаков. В зависимости от соображений предметной области выбирается одна из указанных моделей. Поскольку обе модели трактуются одинаково, в дальнейшем будет рассмотрена только первая из них.

При предположении, что $g(X) = f(X, \Theta)$, где f — известная функция, а Θ — вектор неизвестных параметров, факторизация (4.9) позволяет оценивать $g(X)$ независимо от функции h. Для этого на шкалу возраста наносятся точки $t_{i_1} < \ldots < t_{i_l}$, соответствующие возрасту i_j объекта в момент наступления неисправности, и для каждой точки t_i выписывается $P\{i \mid t_i\}$ — условная вероятность, что среди всех объектов возраста t_i в исследовании неисправность наступит только у i-го объекта при условии, что она действительно наступила у объекта возраста t_i

$$P\{i \mid t_i\} = g(X_i)/\Sigma g(X_j),$$

где суммирование проводится по всем объектам j, в возрасте t_i находившимся в исследовании. Полученные вероятности объединяются в общую функцию условного правдоподобия

$$\ln L = \sum_{i=1}^{l} \ln P\{i_j \mid t_{i_j}\}. \tag{4.10}$$

Параметры Θ оцениваются из условия максимизации $\ln L$. Наиболее часто используется функция $g(X) = \exp\{\Theta' X\}$. Процедуры оценки Θ входят во многие статистические пакеты. Асимптотические свойства Θ изучены пока только в традиционной асимптотике.

4.2.2. Связь между риском и линейной дискриминантной функцией. Формула (4.8) показывает, что всегда возможен

переход от риска события (возникновение неисправности) к вероятности его осуществления за заданный промежуток времени. Проанализируем с этой точки зрения риск

$$r(t,\ X) = g(X)\, h(t) = \exp\{\Theta' X\} \cdot \exp\{e_0 + et\}. \qquad (4.11)$$

Эта формула важна для медицинских приложений, так как $h(t) = \exp\{e_0 + et\}$ достаточно хорошо описывает средний риск кардиоваскулярной смерти для лиц старше 30 лет, а $g(X) = \exp\{\Theta' X\}$ — наиболее часто используемое предположение о $g(X)$.

Пусть H_1, — как прежде, гипотеза, что неисправность не наступила. Если объект был обследован в возрасте s, имел при этом вектор показателей X и пробыл в исследовании T лет, то

$$P\{H_1 | X\} = \exp\left\{-\int_s^{s+T} r(X,\ t)\, \mathrm{d}t\right\} =$$

$$= \exp\{-\exp\{e_0 + \Theta' X + es + \ln((\exp\{eT\} - 1)/e)\}\}. \qquad (4.12)$$

С другой стороны, в классической модели Фишера дискриминантного анализа для описания той же вероятности используется логистическая функция, в которой s — возраст объекта — в момент обследования рассматривается в качестве одной из переменных

$$P\{H_1 | X,\ s\} = \left(1 + \exp\left\{-c_L^{(0)} - c_L' X - c_L^{(p+1)} s\right\}\right)^{-1}. \qquad (4.13)$$

Формулы (4.12) и (4.13) похожи в том смысле, что в обеих в качестве аргумента используются линейные комбинации координат X и s, но они различны аналитически.

Если положить $c_L^{(0)} = -(0{,}3665 + e_0 + \ln((\exp\{eT\} - 1)/e)\,\gamma^{-1}$; $c_L^{(i)} = -\theta^{(i)}\gamma^{-1}$ $(i = 1,\ \ldots,\ p)$; $c_L^{(p+1)} = -e\gamma^{-1}$, то для $\gamma \approx 0{,}80 \div 0{,}90$ оба выражения для вероятности численно близки. Это видно из табл. 4.1, в которой приведены значения функций

$(1 + \exp\{-x\})^{-1}$ и $\exp\{-\exp\{-\gamma x - 0{,}3665\}\}$ для $\gamma = 0{,}80 \div 0{,}95$.

Это позволяет связать оба метода и, в частности, использовать оценки, полученные с помощью дискриминантного анализа, в качестве первого приближения в итеративных процедурах оценки $r(X, t)$.

При работе с риском события информация, содержащаяся в исходных данных, используется более полно, чем при работе с вероятностью осуществления события за время T,

Таблица 4.1

x	$(1+\exp\{-x\})^{-1}$	$\exp\{-\exp\{-\gamma x-0{,}3665\}\}$			
		$\gamma=0{,}80$	$\gamma=0{,}85$	$\gamma=0{,}90$	$\gamma=0{,}95$
0	0,5000	0,5000	0,5000	0,5000	0,5000
0,5	0,6225	0,6284	0,6356	0,6428	0,6498
1,0	0,7311	0,7324	0,7436	0,7544	0,7649
1,25	0,7773	0,7749	0,7870	0,7985	0,8095
1,50	0,8176	0,8116	0,8239	0,8355	0,8464
1,75	0,8520	0,8429	0,8550	0,8663	0,8768
2,0	0,8808	0,8694	0,8811	0,8917	0,9015
2,25	0,9047	0,8917	0,9027	0,9126	0,9215
2,5	0,9241	0,9105	0,9205	0,9295	0,9376
3,0	0,9526	0,9391	0,9473	0,9545	0,9607
3,5	0,9707	0,9587	0,9652	0,9707	0,9754
4,0	0,9820	0,9721	0,9771	0,9812	0,9846

описывается ли она формулой (4.12) или (4.13). Если в факторизации (4.9) $g(X)$ ограничено снизу, а $h(t)$ не убывает с ростом t, то при $T\to\infty$ «разрешающая» сила любого метода ДА стремится к нулю, поскольку все объекты становятся случаями. При использовании функций риска это не страшно, так как при оценке параметров используется информация о том, когда объекты становятся случаями.

4.2.3. Измерение динамики силы влияния факторов. Естественно думать, что влияние того или иного фактора или группы факторов различно в ближайшем и отдаленном периодах. Несмотря на высокую практическую важность количественного изучения динамики силы фактора или интенсивности событий, строго документированные сведения в ряде областей знания практически отсутствуют. Немалую роль в этом сыграло отсутствие до последнего времени подходящего математического аппарата, позволяющего проводить исследование при сравнительно умеренных затратах.

В [271] показано, что повышенное систолическое артериальное давление у мужчины в возрасте 45—60 лет весьма информативно в отношении коронарной смерти в ближайшие 20 месяцев, что со временем информативность падает и что она весьма мала через 90 месяцев после первоначального измерения. Ниже приводятся результаты этой работы с целью демонстрации возможностей, открываемых соответствующим математическим аппаратом.

Пусть s — возраст в момент включения субъекта в исследование, когда проводилось начальное измерение систолического артериального давления, x — величина систолического артериального давления (в мм Hg); $x_{1/4}$ и $x_{3/4}$ —

нижний и верхний квартили распределения x; t — текущий возраст; $r\ (t, s, x)$ — условный риск коронарной смерти для субъекта возраста t при условии, что в возрасте s он имел систолическое артериальное давление x. В исследовании использованы данные из London Busmen Study, эпидемиологического исследования, направленного на выявление риск-

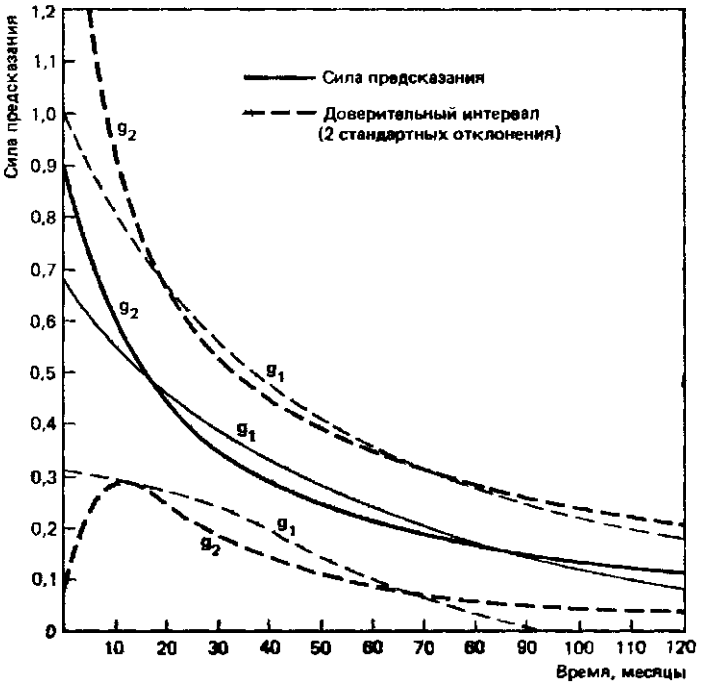


Рис. 4.1. Сила предсказания для двух математических моделей [271]

факторов, ведущих к развитию ишемической болезни сердца. В исследование были включены 684 мужчины в возрасте от 39 до 65 лет. Здоровье каждого из них прослеживалось в течение десяти и более лет. За это время случилось 66 кардиоваскулярных смертей. Если бы имеющиеся данные были разделены на несколько групп согласно возрасту и величине артернального давленія, то численность наблюдений в каждой из получившихся групп была бы недостаточной для каких-либо выводов. Только комплексное использование всего материала на базе предположений о форме зависимо-

сти риска смерти от x, s и t делает анализ возможным. В качестве показателя прогностической силы использовано

$$g = \log_{10}(r(t,\ s,\ x_{3/4})/r(t,\ s,\ x_{1/4})).$$

Модельные предположения о $r(t, s, x)$:

$$r_1(t,\ s,\ x) = \exp\{(a+cx)(1-b)^u\}\cdot h(t); \tag{4.14}$$

$$r_2(t,\ s,\ x) = \exp\{(a+cx)/(1+bu)\}\cdot h(t), \tag{4.15}$$

где a, b, c — неизвестные постоянные; $u = t - s$, а $h(t) = \exp\{e_0 + et\}$, где e_0 и e_1 — постоянные. Анализ можно было бы провести и без конкретизации вида $h(t)$, но при этом на 25 % возросла бы длина доверительных интервалов.

На рис. 4.1 показатель прогностической силы, определенный в предположении (4.14), обозначен g_1, в предположении (4.15) — g_2. Как видим, качественного различия при использовании моделей (4.14) и (4.15) нет. Предсказующая сила убывает очень быстро, уменьшаясь в два раза к концу второго года.

Общая математическая модель для изучения *динамики влияния нескольких факторов* строится [107] из геометрических соображений модели Фишера классического дискриминантного анализа (см. § 2.3). Пусть t, s, X определены как выше, M — вектор средних, а Σ — коварнационная матрица X, тогда

$$r(t;\ s,\ X) = \exp\{(X-M)'\ \Sigma^{-1}\mathbf{D}(u)\,\Theta\}\cdot h(t), \tag{4.16}$$

где

$$\mathbf{D}(u) = \begin{Bmatrix} (1-b_1)^u & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & (1-b_p)^u \end{Bmatrix},\quad u = t-s,\ \Theta,\ b_1,\ \dots,\ b_p — \tag{4.17}$$

неизвестные параметры модели. Асимптотические свойства модели (4.16) в асимптотике растущей размерности пока не исследованы.

4.3. Другие применения ДА

4.3.1. Распознавание сигналов. До сих пор рассматривали задачи, в которых ресурсы, используемые на выработку решающего правила и на саму классификацию нового объекта, не учитывались. При распознавании сигналов картина, как правило, другая: и выработка решающего правила должна быть доступна используемому микропроцессору и классификация (идентификация) объекта должна произойти за конеч-

ное время (часто доли секунды). В качестве примера подобной задачи рассмотрим речевое общение с ЭВМ.

Сначала диктор начитывает ЭВМ используемый им словарный фонд (задает «эталоны»), а затем в ходе общения машина должна правильно идентифицировать произносимые им слова и принимать соответствующие, заранее предусмотренные действия.

В работе [144] образ слова в ЭВМ состоит из $(l \times p)$-матрицы чисел, столбцы которой соответствуют полосам частот в диапазоне от 200 до 5000 Гц и их число фиксировано, строки — последовательным отсчетам времени через 5—15 миллисекунд и их число зависит от длительности произнесения слова, а элементы соответствуют спектральной плотности сигнала на выходе фильтра соответствующей полосы, оцененной за соответствующий интервал времени, и отдельно числа n, показывающего, сколько пересечений нулевого уровня сделано сигналом при произнесении слова.

На повторное произнесение одного и того же слова диктор, вообще говоря, тратит разное время. Поэтому при идентификации слов обязательно производится выравнивание времен так, чтобы допустить небольшие колебания в длительности произношения отдельных звуков.

Отложим по оси абсцисс точки $i = 1, \ldots, m$, соответствующие последовательным отсчетам первого слова, а по оси ординат — точки $j = 1, \ldots, n$, соответствующие отсчетам второго. Рассмотрим далее прямоугольник с вершинами (1, 1), (1, n), (m, n), (m, 1) (рис. 4.2). Точки прямоугольника с целочисленными координатами назовем узлами. Каждой ломаной, выходящей из узла (1,1) и идущей в узел (m, n) по правилу, что из узла (k, l) ломаная может попасть только в один из узлов $(k + 1, l)$, $(k + 1, l + 1)$, $(k, l + 1)$, соответствует вариант сопоставления последовательных отсчетов двух слов. Ограничения на колебания длительности произнесения отдельных звуков можно задать в виде прямых, параллельных диагонали прямоугольника.

Расстояние между двумя словами берется как минимум по разрешенным колебаниям суммы (по столбцам и строкам) квадратов разностей соответствующих элементов матриц, т. е. берется простейшее расстояние. Новое слово идентифицируется с тем эталоном, к которому оно оказывается ближе. В настоящее время на распознавание одного слова при словаре в 100—200 слов и одном дикторе коммерческие системы тратят до 1—2 с времени и обеспечивают среднюю правильность результатов опознания 98—99 %.

Один из путей повышения надежности распознавания — это сопоставление с одним словом нескольких эталонов, как

это предложено в п.1.1.3. При этом возникает чисто статистическая задача выделения небольшого числа представительных вариантов произнесения слова.

4.3.2. Групповая классификация. В технической и реже медицинской диагностике иногда априори известно, что поступившая на диагностику партия из l объектов $\mathbf{X} = (X_1, \ldots, X_l)$ извлечена из одного из классов. При этом наблюдения X_i $(i = 1, \ldots, l)$ при условии, что класс фиксиро-

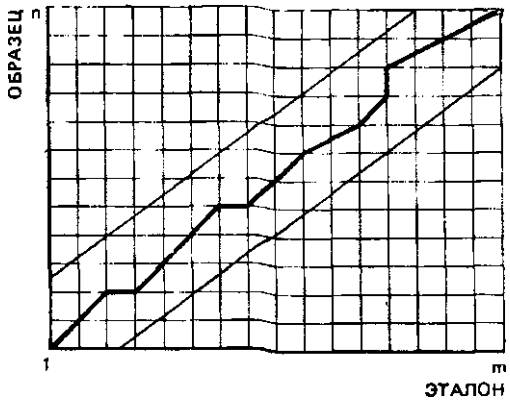


Рис. 4.2. Схема сопоставления координат двух вариантов произнесения одного и того же слова (эталон и образец):
— — траектория сопоставления (ТС);
— — границы отклонения ТС

ван, независимы между собой. Предположим сначала, что нам известны π_j — априорные плотности классов и $f_j(X)$ — плотности распределения X в классах $j = 1, \ldots, k$. Тогда байесовское правило классификации должно быть по аналогии с (1.66) следующим: принимается гипотеза H_i, если $X \in R_i$, где

$$R_i = \left\{ \mathbf{X} : \tilde{\gamma}_{ij} = \sum_{m=1}^{l} \ln \left(f_i(X_m)/f_j(X_m) \right) > - \ln (\pi_i/\pi_j) \right.$$

$$\left. \text{для всех } j \neq i \right\}. \quad (4.18)$$

Очевидно, что если исследователь не знает π_j и f_j, но может оценить их по выборочным данным, то целесообразно в (4.18) заменить π_j и f_j на их оценки.

Частный случай, когда $f_j \in N(M_j, \Sigma)$, изучен в работах [97, 98].

ВЫВОДЫ

1. Одним из основных инструментов применения статистических методов классификации является понятие условной вероятности попадания в один из классов при заданном наблюдении $P\{\text{случай}|X\}$ или, как принято говорить, понятие группы риска. Оно позволяет эффективно выделять объекты, требующие наибольшего внимания, и производить поправку на состав основной и контрольной групп при сравнительных испытаниях.

2. При предсказании будущих событий эффективно введение понятия $r(t, X)$ — мгновенного риска (или интенсивности) стать случаем в момент t при условии, что объект с характеристикой X оставался не-случаем до момента времени t (см. формулу (4.7)). Для того чтобы уменьшить число наблюдений, необходимых для оценки $r(t, X)$, и для более легкой интерпретации $r(t, X)$, Д. Кокс предложил факторизовать риск на два сомножителя $r(t, X) = g(X) \cdot h(t)$ и оценивать параметры, входящие в $g(X)$, независимо от функции $h(t)$. В случае, когда $g(X) = \exp\{X'\Theta\}$ и $h(t) = \exp\{e_0 + et\}$, удается установить связь между подходом с использованием понятия мгновенного риска и подходом с условной вероятностью стать случаем, оцененной с помощью дискриминантного анализа.

3. Понятие мгновенного риска при надлежащей параметризации позволяет изучать динамику изменения (убывания) прогностической силы результатов прошлого обследования объекта с целью определения оптимального интервала между периодическими обследованиями.

4. При распознавании сигналов часто используются простейшие классификационные правила, в которых каждый класс задается набором эталонов, а новый объект приписывается к тому классу, к одному из эталонов которого он оказывается ближе.

5. Если априори известно, что поступившие на классификацию l наблюдений $\mathbf{X} = (X_1, \ldots, X_l)$ являются независимой выборкой из одного из классов, то общее правило классификации строится исходя из плотностей

$$f_j(\mathbf{X}) = \prod_{i=1}^{l} f_j(X_i) \quad (j = 1, \ldots, k).$$

Раздел II. КЛАССИФИКАЦИЯ БЕЗ ОБУЧЕНИЯ: МЕТОДЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ (КЛАСТЕР-АНАЛИЗА) И РАСЩЕПЛЕНИЕ СМЕСЕЙ РАСПРЕДЕЛЕНИЙ

В этом разделе описаны методы классификации объектов (индивидуумов, семей, предприятий, городов, стран, технических систем, признаков и т. д.) O_1, O_2, ..., O_n *в ситуации, когда отсутствуют так называемые обучающие выборки*, а исходная информация о классифицируемых объектах представлена либо в форме матрицы $\mathbf{X}$ «объект — свойство»

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(p)} & x_2^{(p)} & \dots & x_n^{(p)} \end{pmatrix},$$

где $x_i^{(j)}$ — значение j-го признака на i-м статистически обследованном объекте (так что i-й столбец этой матрицы $X_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(p)})'$ характеризует объект O_i, т. е. представляет результат его статистического обследования по всем p анализируемым переменным), либо в форме матрицы $\boldsymbol{\rho}$ попарных взаимных расстояний (близостей) объектов

$$\boldsymbol{\rho} = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \dots & \rho_{nn} \end{pmatrix},$$

где величина ρ_{ij} характеризует взаимную отдаленность (или близость) объектов O_i и O_j.

Переход от формы исходных данных типа «объект — свойство» к форме матрицы попарных расстояний осуществляется посредством задания способа вычисления расстояния (близости) между парой объектов, когда известны координаты (значения признаков) каждого из них (вопросам выбора метрики в исследуемом признаковом пространстве посвящена гл. 11; см. также § 5.2, 7.6).

Обратный переход — от формы записи исходных данных в виде матрицы попарных расстояний (близостей) между объ-

ектами к форме, представленной матрицей «объект — свойство», осуществляется с помощью специального инструментария многомерного статистического анализа, называемого *многомерным метрическим шкалированием* (см. гл. 16).

В зависимости от наличия и характера априорных сведений о природе искомых классов и от конечных прикладных целей исследования следует обратиться либо к гл. 6, где описаны методы *расщепления смесей вероятностных распределений*, которые оказываются полезными в том случае, когда каждый (j-й) класс интерпретируется как параметрически заданная одномодальная генеральная совокупность $f_j(X; \theta_j)$ ($j = 1,2, \ldots, k$) при неизвестном значении определяющего ее векторного значения параметра θ_j и соответственно каждое из классифицируемых наблюдений X_i считается извлеченным из одной из этих (но не известно, из какой именно) генеральных совокупностей; либо к гл. 7, где описаны методы *автоматической классификации* (*кластер-анализа*) многомерных наблюдений, которыми исследователь вынужден пользоваться, когда не имеет оснований для параметрического представления искомых классов, а подчас даже просто для интерпретации классифицируемых наблюдений в качестве выборки из какой-либо вероятностной генеральной совокупности; либо, наконец, к гл. 8, в которой излагаются основные классификационные *процедуры иерархического типа*, используемые в ситуациях, когда «на выходе» исследователь хочет иметь не столько окончательный вариант разбиения анализируемой совокупности объектов на классы, сколько общее наглядное представление о стратификационной структуре этой совокупности (например, в виде специально устроенного графа — *дендрограммы*).

Глава 5. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ, ИСПОЛЬЗУЕМЫЕ В МЕТОДАХ КЛАССИФИКАЦИИ БЕЗ ОБУЧЕНИЯ

5.1. Общая (нестрогая) постановка задачи классификации объектов или признаков в условиях отсутствия обучающих выборок

Говоря о классификации совокупности *объектов*, подразумеваем, что каждый из них задан соответствующим *столбцом* матрицы **X** либо геометрическая структура их попарных расстояний (близостей) задана матрицей **ρ**. Аналогич-

но интерпретируется исходная информация в задаче классификации совокупности *признаков*, с той лишь разницей, что каждый из признаков задается соответствующей *строкой* матрицы **X**. В дальнейшем, если это специально не оговорено, не будем разделять изложение этой проблемы на «объекты» и «признаки», поскольку все постановки задач и основная методологическая схема исследования здесь общие.

В общей (нестрогой) постановке *проблема классификации объектов заключается в том, чтобы всю анализируемую совокупность объектов* $\mathbf{O} = \{O_i\}$ $(i = \overline{1, n})$, *статистически представленную в виде матриц* **X** или **ρ**, *разбить на сравнительно небольшое число (заранее известное или нет) однородных, в определенном смысле, групп или классов.*

Для формализации этой проблемы удобно интерпретировать анализируемые объекты в качестве точек в соответствующем признаковом пространстве. Если исходные данные представлены в форме матрицы (**X**), то эти точки являются непосредственным геометрическим изображением многомерных наблюдений $X_1, X_2, \ldots, X_n$ в p-мерном пространстве $\mathbf{\Pi}^p(X)$ с координатными осями $0x^{(1)}, 0x^{(2)}, \ldots, Ox^{(p)}$. Если же исходные данные представлены в форме матрицы попарных взаимных расстояний **ρ**, то исследователю не известны непосредственно координаты этих точек, но зато задана структура попарных расстояний (близостей) между объектами. Естественно предположить, что геометрическая близость двух или нескольких точек в этом пространстве означает близость «физических» состояний соответствующих объектов, их однородность. Тогда проблема классификации состоит в разбиении анализируемой совокупности точек — наблюдений на сравнительно небольшое число (заранее известное или нет) классов таким образом, чтобы объекты, принадлежащие одному классу, находились бы на сравнительно небольших расстояниях друг от друга. Полученные в результате разбиения классы часто называют *кластерами* (таксонами, образами) [1], а методы их нахождения соответственно кластер-анализом, численной таксономией, распознаванием образов с самообучением.

Однако, берясь за решение задачи классификации, исследователь с самого начала должен четко представлять,

---

[1] *Cluster* (англ.) — гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством. *Taxon* (англ.) — систематизированная группа любой категории (термин биологического происхождения). Название «кластер-анализ» для совокупности методов решения задач такого типа было впервые введено, по-видимому, Трайоном в 1939 г. (см.: Tryon R. C. Cluster Analysis// Ann. Arb., Edw. Brothers. — 1939).

какую именно из двух задач он решает. Рассматривает ли он обычную задачу разбиения статистически обследованного (p-мерного) диапазона изменения значений анализируемых признаков на *интервалы* (*гиперобласти*) *группирования*, в результате решения которой исследуемая совокупность объектов разбивается на некоторое число групп так, что объекты такой одной группы находятся друг от друга на сравнительно небольшом расстоянии (многомерный аналог задачи построения интервалов группирования при обработке одномерных наблюдений). Либо он пытается определить *естественное расслоение* исходных наблюдений на четко выраженные кластеры, сгустки, лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные части. В вероятностной интерпретации (т. е. если интерпретировать классифицируемые наблюдения X_1, X_2, ..., X_n как выборку из некоторой многомерной генеральной совокупности, описываемой функцией плотности или полигоном распределения $f(X)$, как правило, не известными исследователю) вторая задача может быть сформулирована как задача выявления областей повышенной плотности наблюдений, т. е. таких областей возможных значений анализируемого многомерного признака X, *которые соответствуют локальным максимумам функции* $f(X)$.

Если первая задача — задача построения областей группирования — всегда имеет решение, то при второй постановке результат может быть и отрицательным: может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры (например, образует один общий кластер).

Из методологических соображений (в частности, для упрощения понимания читателем некоторых основных идей теории автоматической классификации и для создания удобной схемы исследования свойств различных классификационных процедур) будем иногда вводить в рассмотрение теоретические вероятностные характеристики анализируемой совокупности: генеральную совокупность, плотность (полигон) распределения или соответствующую вероятностную меру $P(dX)$, теоретические средние значения, дисперсии, ковариации и т. п. Очевидно, если мысленно «продолжить» множество классифицируемых наблюдений *до всей генеральной совокупности* (методологический прием, уже использованный в гл. 1), задача классификации заключается в разбиении анализируемого признакового пространства $\mathbf{\Pi}^p(X)$ на некоторое число непересекающихся областей. Условимся в дальнейшем называть такую схему *теоретико-вероятностной модификацией задачи кластер-анализа*.

5.2. Расстояния между отдельными объектами и меры близости объектов друг к другу

Наиболее трудиым и наименее формализованным в задаче автоматической классификации является момент, связанный с определением понятия однородности объектов.

В общем случае понятие однородности объектов определяется заданием правила вычисления величины ρ_{ij}, характеризующей либо расстояние $d(O_i, O_j)$ между объектами O_i и O_j из исследуемой совокупности $O(i, j = 1,2,\ldots, n)$, либо степень близости (сходства) $r(O_i, O_j)$ тех же объектов. Если задана функция $d(O_i, O_j)$, то близкие в смысле этой метрики объекты считаются однородными, принадлежащими к одному классу. Естественно, при этом необходимо сопоставление $d(O_i, O_j)$ с некоторым пороговым значением, определяемым в каждом конкретном случае по-своему.

Аналогично используется для формирования однородных классов и упомянутая выше мера близости $r(O_i, O_j)$, при задании которой иужио помнить о необходимости соблюдения следующих естественных требованиий: требования симметрии $(r(O_i, O_j) = r(O_j, O_i))$; требования максимального сходства объекта с самим собой $(r(O_i, O_i) = \max r(O_i, O_j))$ и требования при заданной метрике монотонного убывания $r(O_i, O_j)$ по $d(O_i, O_j)$, т. е. из $d(O_k, O_l) \geqslant d(O_i, O_j)$ должно с необходимостью следовать выполнение неравенства $r(O_k, O_l) \leqslant r(O_i, O_j)$.

Конечно, *выбор метрики (или меры близости) является узловым моментом исследования*, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при заданном алгоритме разбиения. В каждой конкретной задаче этот выбор должен производиться по-своему. При этом решение данного вопроса зависит в основном от главных целей исследования, физической и статистической природы вектора наблюдений X, полноты апрнорных сведений о характере вероятностного распределения X. Так, например, если нз конечных целей исследования и из прнроды вектора X следует, что понятие однородиой группы естественно интерпретировать как генеральиую совокупность с одновершинной плотностью (полигоном частот) распределения, и если к тому же известен общий вид этой плотности, то следует воспользоваться общим подходом, описанным в гл. 6. Если, кроме того, известно, что наблюдения X_i извлекаются из *нормальных* генеральных совокупностей с *одной и той же матрицей ковариаций*, то естествеиной мерой отдаленности двух объектов друг от друга является *расстояние махаланобисского типа* (см. ниже).

В качестве примеров расстояний и мер близости, сравнительно широко используемых в задачах кластер-анализа, приведем здесь следующие.

Общий вид метрики махаланобисского типа. В общем случае зависимых компонент $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ вектора наблюдении X и их различной значимости в решении вопроса об отнесении объекта (наблюдения) к тому или иному классу обычно пользуются *обобщенным* («*взвешенным*») *расстоянием махаланобисского типа*, задаваемым формулой [1]

$$d_{\Phi}(X_i, X_j) = \sqrt{(X_i - X_j)' \Lambda' \Sigma^{-1} \Lambda (X_i - X_j)}.$$

Здесь Σ — ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения X_i, а Λ — некоторая симметричная неотрицательно-определенная матрица «весовых» коэффициентов $\lambda_{m\,q}$, которая чаще всего выбирается диагональной [195, 279].

Следующие три вида расстояний, хотя и являются частными случаями метрики d_{Φ}, все же заслуживают специального описания.

Обычное евклидово расстояние

$$d_{\mathrm{E}}(X_i, X_j) = \sqrt{(x_i^{(1)} - x_j^{(1)})^2 + (x_i^{(2)} - x_j^{(2)})^2 + \ldots + (x_i^{(p)} - x_j^{(p)})^2}.$$

К ситуациям, в которых использование этого расстояния можно признать оправданным, прежде всего относят следующие:

наблюдения X извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с ковариационной матрицей вида $\sigma^2 \cdot \mathrm{I}$, т. е. компоненты X взаимно независимы и имеют одну и ту же дисперсию;

компоненты $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ вектора наблюдении X однородны по своему физическому смыслу, причем установлено, например с помощью опроса экспертов, что все они одинаково важны с точки зрения решения вопроса об отнесении объекта к тому или иному классу;

признаковое пространство совпадает с геометрическим пространством нашего бытия, что может быть лишь в случаях $p = 1, 2, 3$, и понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве, например классификация попаданий при стрельбе по цели.

---

[1] В случаях, когда каждый объект O_i представлен вектором признаков X_i (т. е. в случае исходных данных, представленных в форме X), часто удобнее в формулах и различных соотношениях вместо O_i писать сразу X_i. Например, $d(X_i, X_j)$ вместо $d(O_i, O_j)$.

«Взвешенное» евклидово расстояние

$$d_{\text{вЕ}}(X_i,\ X_j) =$$

$$= \sqrt{\omega_1 (x_i^{(1)} - x_j^{(1)})^2 + \omega_2 (x_i^{(2)} - x_j^{(2)})^2 + \ldots + \omega_p (x_i^{(p)} - x_j^{(p)})^2}.$$

Обычно применяется в ситуациях, в которых так или иначе удается приписать каждой из компонент $x^{(k)}$ вектора наблюдений X некоторый неотрицательный «вес» ω_k, пропорциональный степени его важности с точки зрения решения вопроса об отнесении заданного объекта к тому или иному классу. Удобно полагать при этом $0 \leqslant \omega_k \leqslant 1$, $k = \overline{1, p}$.

Определение весов ω_k связано, как правило, с дополнительным исследованием, например получением и использованием обучающих выборок, организацией опроса экспертов и обработкой их мнений, использованием некоторых специальных моделей. Попытки определения весов ω_k только по информации, содержащейся в исходных данных [72, 330], как правило, не дают желаемого эффекта, а иногда могут лишь отдалить от истинного решения. Достаточно заметить, что в зависимости от весьма тонких и незначительных вариаций физической и статистической природы исходных данных можно привести одинаково убедительные доводы в пользу двух диаметрально противоположных решений этого вопроса: выбирать ω_k пропорционально величине среднеквадратической ошибки признака $x^{(k)}$ [138] либо пропорционально обратной величине среднеквадратической ошибки этого же признака [332, 72, 330].

Хеммингово расстояние. Используется как мера различия объектов, задаваемых дихотомическими признаками. Оно задается с помощью формулы

$$d_H(X_i,\ X_j) = \sum_{s=1}^{p} \left| x_i^{(s)} - x_j^{(s)} \right|$$

и, следовательно, равно числу ν_{ij} несовпадений значений соответствующих признаков в рассматриваемых i-м и j-м объектах.

Другие меры близости для дихотомических признаков. Меры близости объектов, описываемых набором дихотомических признаков, обычно основаны на характеристиках $\nu_{ij}^{(0)}$, $\nu_{ij}^{(1)}$ и $\nu_{ij} = \nu_{ij}^{(0)} + \nu_{ij}^{(1)}$, где $\nu_{ij}^{(0)}$ ($\nu_{ij}^{(1)}$) — число нулевых (единичных) компонент, совпавших в объектах X_i и X_j. Так, например, если из каких-либо профессиональных соображений или априорных сведений следует, что все p признаков исследуемых объектов можно считать равноправными, а эффект от совпадения или несовпадения нулей такой

же, что и от совпадения или несовпадения единиц, то в качестве меры близости объектов X_i и X_j используют величину $r(X_i, X_j) = \frac{\nu_{ij}}{p}$.

Весьма полный обзор различных мер близости объектов, описываемых дихотомическими признаками, читатель найдет в [136, 29].

Меры близости и расстояния, задаваемые с помощью потенциальной функции. Во многих задачах математической статистики, теории вероятностей, физической теории потенциала и теории распознавания образов, или классификации многомерных наблюдений, оказываются полезными некоторые специально устроенные функции $K(X, Y)$ от двух векторных переменных X и Y, а чаще всего просто от расстояния $d(X, Y)$ между этими переменными, которые будем называть *потенциальными* [1].

Так, например, если пространство $\Pi^p(X)$ всех мыслимых значений исследуемого вектора X разбито на полную систему непересекающихся односвязных компактных множеств или однородных классов $S_1, \ldots, S_k$ и потенциальная функция $K(X, Y)$ определена для $X \in \Pi^p(X)$ и $Y \in \Pi^p(X)$ следующим образом:

$$K(X, Y) = \begin{cases} 1, \text{ если } X \in S_j,\ Y \in S_j\ (j = 1, 2, \ldots, k), \\ 0 \text{ в противном случае,} \end{cases}$$

то с помощью этой функции удобно строить обычные эмпирические гистограммы (оценки плотности распределения $\widehat{f_n}(U)$) по имеющимся наблюдениям $X_1, X_2, \ldots, X_n$. Действительно, легко видеть, что

$$\widehat{f}_n(U) = \frac{1}{W(S_{j(U)}) \cdot n} \sum_{i=1}^{n} K(U, X_i) = \frac{\nu(U)}{nW(S_{j(U)})}, \qquad (5.1)$$

где $\nu(U)$ — число наблюдений, попавших в класс $S_{j(U)}$, содержащий точку U, а $W(S_{j(U)})$ — объем области $S_{j(U)}$ (геометрическая интерпретация для одномерного случая показана на рис. 5.1).

Если в исследуемом факторном пространстве $\Pi^p(X)$ задана метрика $d(U, V)$, то можно не связывать себя за-

[1] В некоторых работах можно встретить, по существу, те же функции, но под другим названием, например window — «окно» [280, 290]. Определение «потенциальные функции» обосновывается тем, что примером подобных зависимостей в физике является потенциал, определенный для любой точки пространства, но зависящий от того, где расположен источник потенциала.

ранее зафиксированным разбиением $\Pi^p(X)$ на классы, а задавать $K(U, V)$ как монотонно убывающую функцию расстояния $d(U, V)$. Например,

$$K(U, V) = e^{-\alpha d^2(U, V)}, \ \alpha > 0;$$

$$K(U, V) = [1 + \alpha d^2(U, V)]^{-1}, \ \alpha > 0. \qquad (5.2)$$

Приведем здесь еще лишь одну достаточно общую форму связи между $d(U, V)$ и $K(U, V)$, в которой расстояние d

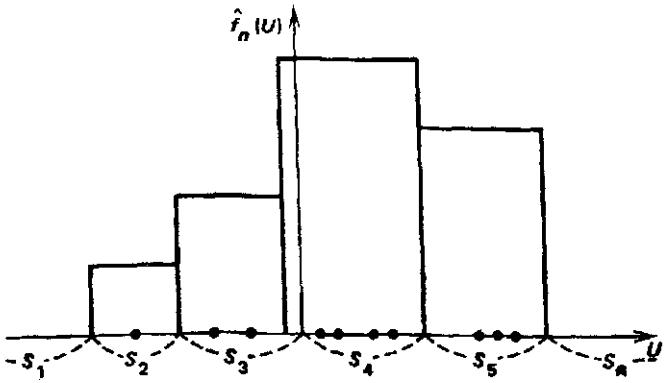


Рис. 5.1. Гистограмма $\widehat{f_n}(U)$, построенная с помощью разбиения на группы выборочной одномерной совокупности $X_1, \ldots, X_n$

выступает как функция некоторых значений потенциальной функции K:

$$d(U, V) = \sqrt{K(U, U) + K(V, V) - 2K(U, V)}. \qquad (5.3)$$

В частности, выбрав в качестве $K(U, V)$ скалярное произведение векторов U и V, т. е. положив

$$K(U, V) = (U, V) = \sum_{i=1}^{p} u^{(i)} v^{(i)},$$

получим по формуле (5.3) обычное евклидово расстояние d_E.

Легко понять, что и в случае задания потенциальной функции в виде соотношений (5.2) формулы (5.1) позволяют строить статистические оценки плотности распределения (5.1), хотя график функции $\widehat{f_n}(U)$ будет уже не ступенчатым, а сглаженным. При отсутствии метрики в пространстве $\Pi^p(X)$ функции $K(U, V)$ могут быть использованы в качестве меры близости объектов U и V, а также объектов и це-

лых классов и классов между собой. В первом случае эта мера позволяла получить лишь качественный ответ: объекты близки, если U и V принадлежат одному классу, и объекты далеки — в противном случае; в двух других случаях мера близости является количественной характеристикой.

О физически содержательных мерах близости объектов. В некоторых задачах классификации объектов, не обязательно описываемых количественно, естественнее использовать в качестве меры близости объектов (или расстояния между ними) некоторые физически содержательные числовые параметры, так или иначе характеризующие взаимоотношения между объектами. Примером может служить задача классификации с целью агрегирования отраслей народного хозяйства, решаемая на основе матрицы межотраслевого баланса [97]. Таким образом, классифицируемым объектом в данном примере является отрасль народного хозяйства, а матрица межотраслевого баланса представлена элементами s_{ij}, где под s_{ij} подразумевается сумма годовых поставок в денежном выражении i-й отрасли в j-ю. В качестве матрицы близости $\{r_{ij}\}$ в этом случае естественно взять, например, симметризованную нормированную матрицу межотраслевого баланса. При этом под нормировкой понимается преобразование, при котором денежное выражение поставок из i-й отрасли в j-ю заменяется долей этих поставок по отношению ко всем поставкам i-й отрасли. Симметризацию же нормированной матрицы межотраслевого баланса можно проводить различными способами. Так, например, в [97] близость между i-й и j-й отраслями выражается либо через среднее значение их взаимных нормированных поставок, либо через комбинацию из их взаимных нормированных поставок.

О мерах близости числовых признаков (отдельных факторов). Решение задач классификации многомерных данных, как правило, предусматривает в качестве предварительного этапа исследования реализацию методов, позволяющих существенно сократить размерность исходного факторного пространства, выбрать из компонент $x^{(1)}, \ldots, x^{(p)}$ наблюдаемых векторов X сравнительно небольшое число наиболее существенных, наиболее информативных. Для этих целей бывает полезно рассмотреть каждую из компонент $x^{(1)}, \ldots, x^{(p)}$ в качестве объекта, подлежащего классификации. Дело в том, что разбиение признаков $x^{(1)}, \ldots, x^{(p)}$ на небольшое число однородных в некотором смысле групп позволит исследователю сделать вывод, что компоненты, входящие в одну группу, в определенном смысле сильно связаны друг с другом и несут информацию о каком-то одном свойстве ис-

следуемого объекта. Следовательно, можно надеяться, что не будет большого ущерба в информации, если для дальнейшего исследования оставим лишь по одному представителю от каждой такой группы.

Чаще всего в подобных ситуациях в качестве мер близости между отдельными признаками $x^{(i)}$ и $x^{(j)}$, так же как и между наборами таких признаков, используются различные характеристики степени их коррелированности и в первую очередь коэффициенты корреляции. Проблеме сокращения размерности анализируемого признакового пространства специально посвящен раздел III книги. Более подробно вопросы построения и использования расстояний и мер близости между отдельными объектами рассмотрены в [136, 288, 296, 29].

5.3. Расстояние между классами и мера близости классов

При конструировании различных процедур классификации (кластер-процедур) в ряде ситуаций оказывается целесообразным введение понятия расстояния между целыми группами объектов, так же как и понятия меры близости двух групп объектов. Приведем примеры наиболее распространенных расстояний и мер близости, характеризующих взаимное расположение отдельных групп объектов. Пусть S_i — i-я группа (класс, кластер) объектов, n_j — число объектов, образующих группу S_i, вектор $\overline{X}(i)$ — среднее арифметическое векторных наблюдений, входящих в S_i (другими словами, $\overline{X}(i)$ — «центр тяжести» i-й группы), а $\rho(S_l, S_m)$ — расстояние между группами S_l и S_m.

Ниже приводятся наиболее употребительные и наиболее общие расстояния и меры близости между классами объектов.

Расстояние, измеряемое по принципу «ближнего соседа» («nearest neighbour»)

$$\rho_{\min}(S_l, S_m) = \min_{X_i \in S_l,\ X_j \in S_m} d(X_i, X_j). \tag{5.4}$$

Расстояние, измеряемое по принципу «дальнего соседа» («furthest neighbour») [262, 224]

$$\rho_{\max}(S_l, S_m) = \max_{X_i \in S_l,\ X_j \in S_m} d(X_i, X_j). \tag{5.5}$$

Расстояние, измеряемое по «центрам тяжести» групп [262, 224]

$$\rho(S_l, S_m) = d\left(\overline{X}(l), \overline{X}(m)\right). \tag{5.6}$$

Мера близости групп, основанная на потенциальной функции [57]

$$r(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} K(X_i, X_j).$$

Расстояние, измеряемое по принципу «средней связи». Определяется [262, 224] как арифметическое среднее всевозможных попарных расстояний между представителями рассматриваемых групп, т. е.

$$\rho_{cp}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d(X_i, X_j). \qquad (5.7)$$

Естественно задать вопрос: нельзя ли получить достаточно общую формулу, определяющую расстояние между классами по заданному расстоянию между отдельными элементами (наблюдениями), которая включила бы в себя в качестве частных случаев все рассмотренные выше виды расстояний?

Изящное обобщение такого рода, основанное на понятии так называемого «обобщенного среднего», а точнее, степенного среднего, было предложено А. Н. Колмогоровым. Под обобщенным средним величин $c_1, c_2, \ldots, c_N$ понимается выражение вида $M^F(c_1, c_2, \ldots, c_N) = F^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} F(c_i)\right)$, в котором $F(u)$ — некоторая функция и соответственно F^{-1} — преобразование, обратное к F. Частным случаем обобщенного среднего является *степенное среднее*, определяемое как

$$M_\tau(c_1, c_2, \ldots, c_N) = \left(\frac{1}{N}\sum_{i=1}^{N} c_i^\tau\right)^{\frac{1}{\tau}}.$$

Нетрудно показать, что (при $c_i > 0$, $i = 1, 2, \ldots, N$)

$$M_{-\infty}(c_1, c_2, \ldots, c_N) = \min_{1 \leqslant i \leqslant N}(c_i);$$

$$M_{+\infty}(c_1, c_2, \ldots, c_N) = \max_{1 \leqslant i \leqslant N}(c_i);$$

$$M_0(c_1, c_2, \ldots, c_N) = \left(\prod_{i=1}^{N} c_i\right)^{1/N} \text{— геометрическое среднее;}$$

$$M_1(c_1, c_2, \ldots, c_N) = \frac{1}{N}\sum_{i=1}^{N} c_i \text{— арифметическое среднее.}$$

Обобщенное (по Колмогорову) расстояние между классами, или обобщенное K-расстояние, вычисляется по формуле

$$\rho_{\tau}^{(K)}(S_l S_m) = \left[\frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d^{\tau}(X_i, X_j)\right]^{\frac{1}{\tau}}. \qquad (5.8)$$

В частности, при $\tau \to \infty$ и при $\tau \to -\infty$ имеем

$$\rho_{\infty}^{(K)}(S_l, S_m) = \rho_{\max}(S_l, S_m); \quad \rho_{-\infty}^{(K)}(S_l, S_m) = \rho_{\min}(S_l, S_m).$$

Очевидно также

$$\rho_{1}^{(K)}(S_l, S_m) = \rho_{\text{ср}}(S_l, S_m).$$

Из (5.8) следует, что если $S(m, q) = S_m \cup S_q$ — группа элементов, полученная путем объединения кластеров S_m и S_q, то обобщенное K-расстояние между кластерами S_l и $S(m, q)$ определяется формулой

$$\rho_{\tau}^{(K)}(S_l, S(m,q)) = \left\{\frac{n_m \left[\rho_{\tau}^{(K)}(S_l, S_m)\right]^{\tau} + n_q \left[\rho_{\tau}^{(K)}(S_l, S_q)\right]^{\tau}}{n_m + n_q}\right\}^{\frac{1}{\tau}}$$

Понятие расстояния между группами элементов особенно важно в так называемых *агломеративных иерархических кластер-процедурах*, поскольку принцип работы таких алгоритмов состоит в последовательном объединении элементов, а затем и целых групп, сначала самых близких, а потом все более и более отдаленных друг от друга. Подробнее об агломеративных иерархических процедурах см. в гл. 8. Учитывая специфику подобных процедур, для задания расстояния между классами оказывается достаточным определить порядок пересчета расстояния между классом S_l и классом $S(m, q) = S_m \cup S_q$, являющимся объединением двух других классов S_m и S_q, по расстояниям $\rho_{lm} = \rho(S_l, S_m)$, $\rho_{lq} = \rho(S_l, S_q)$ и $\rho_{mq} = \rho(S_m, S_q)$ между этими классами. В [255] предлагается следующая общая формула для вычисления расстояния между некоторым классом S_l и классом $S(m, q)$:

$$\rho_{l(m,q)} = \rho(S_l, S(m, q)) = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta\,|\rho_{lm} - \rho_{lq}|, \qquad (5.9)$$

где α, β, γ и δ — числовые коэффициенты, значения которых и определяют специфику процедуры, ее нацеленность на решение той или иной экстремальной задачи. Так, например, полагая $\alpha = \beta = -\delta = \frac{1}{2}$ и $\gamma = 0$, приходим к расстоянию, измеряемому по принципу «ближайшего соседа». Если

же положить $\alpha = \beta = \delta = \frac{1}{2}$ и $\gamma = 0$, то расстояние между двумя классами определится как расстояние между двумя самыми далекими элементами этих классов, по принципу «дальнего соседа». И наконец, выбор коэффициентов соотношения (5.9) по формулам

$$\alpha = \frac{n_m}{n_m + n_q}, \quad \beta = \frac{n_q}{n_m + n_q}, \quad \gamma = \delta = 0$$

приводит к расстоянию $\rho_{\text{ср}}$ между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой — из другого.

То, что формула для $\rho_{l(m,\ q)}$ и, в частности, выбор коэффициентов α, β, γ и δ в этой формуле зачастую определяют нацеленность соответствующей агломеративной иерархической процедуры на решение той или иной экстремальной задачи, т. е. в каком-то смысле определяет ее оптимальную критерийную установку, поясняет, например, следующий результат [331]. Оказывается, если для вычисления $\rho_{l\ (m,\ q)}$ воспользоваться следующей модификацией формулы (5.9):

$$\rho^2_{l\ (m,\ q)} = \frac{n_l + n_m}{n_l + n_m + n_q}\rho^2_{lm} + \frac{n_l + n_q}{n_l + n_m + n_q}\rho^2_{lq} - \frac{n_l}{n_l + n_m + n_q}\rho^2_{mq}, \tag{5.10}$$

то соответствующий агломеративный иерархический алгоритм обладает тем свойством, что на каждом шаге объединение двух классов приводит к минимальному увеличению общей суммы квадратов расстояний между элементами внутри классов. Отметим сразу, что такая пошаговая оптимальность алгоритма в указанном смысле, вообще говоря, не влечет его оптимальности в том же смысле для любого наперед заданного числа классов, на которые требуется разбить исходную совокупность элементов.

5.4. Функционалы качества разбиения на классы и экстремальная постановка задачи кластер-анализа. Связь с теорией статистического оценивания параметров

Естественно попытаться определить сравнительное качество различных способов разбиения заданной совокупности элементов на классы, т. е. определить тот количественный критерий, следуя которому можно было бы предпочесть од-

ию разбиение другому. С этой целью в постановку задачи кластер-анализа часто вводится понятие так называемого *функционала качества разбиения* $Q(S)$, определенного на множестве всех возможных разбиений. Функционалом он называется потому, что чаще всего разбиение S задается, вообще говоря, набором дискриминантных функций $\delta_1(X)$, $\delta_2(X)$, ... Тогда под наилучшим разбиением S^* понимается то разбиение, на котором достигается экстремум выбранного функционала качества. Выбор того или иного функционала качества, как правило, осуществляется весьма произвольно и опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо строгую формализованную систему.

Приведем примеры наиболее распространенных функционалов качества разбиения и попытаемся обосновать выбор некоторых из них в рамках одной из моделей статистического оценивания параметров.

5.4.1. Функционалы качества разбиения при заданном числе классов. Пусть исследователем уже выбрана метрика d в пространстве $\Pi^p(X)$ и пусть $S=(S_1, S_2, \ldots, S_k)$ — некоторое фиксированное разбиение наблюдений $X_1, X_2, \ldots, X_n$ на заданное число k классов $S_1, S_2, \ldots, S_k$.

За функционалы качества часто берутся следующие характеристики:

сумма («взвешенная») внутриклассовых дисперсий

$$Q_1(S)=\sum_{l=1}^{k}\sum_{X_i\in S_l} d^2(X_i, \overline{X}(l)), \qquad (5.11)$$

весьма широко используется в задачах кластер-анализа в качестве критерийной оценки разбиения [268];

сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S)=\sum_{l=1}^{k}\sum_{X_i, X_j\in S_l} d^2(X_i, X_j)$$

либо

$$Q_2'(S)=\sum_{l=1}^{k}\frac{1}{n_l}\sum_{X_i, X_j\in S_l} d^2(X_i, X_j),$$

в большинстве ситуаций приводит к тем же наилучшим разбиениям, что и $Q_1(S)$, и тоже используется для сравнения кластер-процедур [228];

обобщенная внутриклассовая дисперсия $Q_3(S)$ является, как известно [16, с. 231], одной из характеристик степени

рассеивания многомерных наблюдений одного класса (генеральной совокупности) около своего «центра тяжести». Следуя обычным правилам вычисления выборочной ковариационной матрицы $\mathbf{W}_l$, отдельно по наблюдениям, попавшим в какой-то один класс S_l, получаем

$$Q_3(S) = \det\left(\sum_{l=1}^{k} n_l \mathbf{W}_l\right), \tag{5.12}$$

где под $\det \mathbf{A}$ понимается «определитель матрицы $\mathbf{A}$», а элементы $w_{qm}(l)$ выборочной ковариационной матрицы $\mathbf{W}_l$ класса S_l подсчитываются по формуле

$$w_{qm}(l) = \frac{1}{n_l} \sum_{X_i \in S_l} \left(x_i^{(q)} - \bar{x}^{(q)}(l)\right)\left(x_i^{(m)} - \bar{x}^{(m)}(l)\right), \quad q, m = 1, 2, \ldots, p, \tag{5.13}$$

где $x_i^{(v)}$ — v-я компонента многомерного наблюдения X_i, а $\bar{x}^{(v)}(l)$ — среднее значение v-й компоненты, подсчитанное по наблюдениям l-го класса.

Встречается и другой вариант использования понятия обобщенной дисперсии как характеристики качества разбиения, в котором операция суммирования $\mathbf{W}_l$ по классам заменена операцией умножения

$$Q_4(S) = \prod_{l=1}^{k} (\det \mathbf{W}_l)^{n_l}.$$

Как видно из формул (5.12) и (5.13), функционал $Q_3(S)$ является средней арифметической (по всем классам) характеристикой обобщенной внутриклассовой дисперсии, в то время как функционал $Q_4(S)$ пропорционален средней геометрической характеристике тех же величин.

Использование функционалов $Q_3(S)$ и $Q_4(S)$ является особенно уместным в ситуациях, при которых исследователь в первую очередь задается вопросом: не сосредоточены ли наблюдения, разбитые на классы $S_1, S_2, \ldots, S_k$, в пространстве размерности, меньшей, чем p?

Замечание. При теоретико-вероятностной модификации схем кластер-анализа соответственно видоизменится запись приведенных выше функционалов. Так, например,

$$Q_1'(S) = \sum_{l=1}^{k} \int_{S_l} d(X, \bar{X}(l))\, P(\mathrm{d}X),$$

где

$$\overline{X}(l) = \frac{1}{P(S_l)} \int\limits_{S_l} XP(\mathrm{d}X)$$

или

$$Q_2'(S) = \sum_{l=1}^{k} \frac{1}{P(S_l)} \int\limits_{S_l} \int\limits_{S_l} d^2(X, Y) P(\mathrm{d}X) P(\mathrm{d}Y). \qquad (5.14)$$

5.4.2. Функционалы качества разбиения при неизвестном числе классов. В ситуациях, когда исследователю заранее не известно, на какое число классов подразделяются исходные многомерные наблюдения $X_1, X_2, \ldots, X_n$, функционалы качества разбиения $Q(S)$ выбирают чаще всего в виде простой алгебраической комбинации (суммы, разности, произведения, отношения) двух функционалов $I_1(S)$ и $I_2(S)$, один из которых I_1 является убывающей (невозрастающей) функцией числа классов k и характеризует, как правило, внутриклассовый разброс наблюдений, а второй I_2 — возрастающей (неубывающей) функцией числа классов k. При этом интерпретация функционала I_2 может быть различной. Под I_2 понимается иногда и некоторая мера взаимной удаленности (близости) классов, и мера тех потерь, которые приходится нести исследователю при излишней детализации рассматриваемого массива исходных наблюдений, и величина, обратная так называемой «мере концентрации» всей структуры точек, полученной при разбиении исследуемого множества наблюдений на k классов. В [218], например, предлагается брать

$$I_1(S) = \sum_{l=1}^{k} \sum_{X_i \in S_l} d(X_i, \overline{X}(l))$$

и

$$I_2(S) = ck(S),$$

где $k(S)$ — число классов, получающихся при разбиении S, а c — некоторая положительная постоянная, характеризующая потери исследователя при увеличении числа классов на единицу.

Другой вариант функционалов качества такого типа можно найти, например, в [57], где полагают

$$I_1'(S) = \frac{1}{k} \sum_{l=1}^{k} \left\{ \frac{2}{n_l (n_l - 1)} \right\} \sum_{X_i \in S_l,\ X_j \in S_l} K(X_i,\ X_j);$$

$$I_2'(S) = -\frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j>i}^{k} r(S_i,\ S_j).$$

Здесь $K(X, Y)$ — упомянутая выше потенциальная функция, а $r(S_i, S_j)$ — мера близости i-го и j-го классов, основанная на потенциальной функции (5.6).

Очевидно, в первом случае будем искать разбиение S^*, минимизирующее значение функционала

$$Q(S) = I_1(S) + I_2(S), \tag{5.15}$$

в то время как во втором случае требуется найти разбиение S^0, которое максимизировало бы значение функционала

$$Q'(S) = I_1'(S) + I_2'(S). \tag{5.16}$$

Весьма гибким и достаточно общим подходом, реализующим идею одновременного учета двух функционалов, является подход, основанный на схеме, предложенной А.Н. Колмогоровым. Эта схема опирается на понятия *меры концентрации* $Z_\tau(S)$ точек, соответствующей разбиению S, и *средней меры внутриклассового рассеяния* $I_\tau^{(K)}(S)$, характеризующей то же разбиение S.

Под мерой концентрации $Z_\tau(S)$ предлагается понимать степенное среднее (см. § 5.3) вида

$$Z_\tau(S) = M_\tau\left(\frac{\nu(X_1)}{n},\ \frac{\nu(X_2)}{n},\ \ldots,\ \frac{\nu(X_n)}{n}\right) =$$

$$= \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\nu(X_i)}{n}\right)^\tau\right]^{\frac{1}{\tau}}, \tag{5.17}$$

где $\nu(X_i)$ — число элементов в кластере, содержащем точку X_i, а выбор числового параметра τ находится в распоряжении исследователя и зависит от конкретных целей разбиения. При выборе τ полезно иметь в виду следующие частные случаи $Z_\tau(S)$:

$$Z_{-1}(S) = \frac{1}{k},$$

где k — число различных кластеров в разбиении S; $\log Z_0(S) = \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$ — естественная информационная мера концентрации;

$$Z_\infty(S) = \max_{1 \leqslant i \leqslant k} \left(\frac{n_i}{n} \right);$$

$$Z_{-\infty}(S) = \min_{1 \leqslant i \leqslant k} \left(\frac{n_i}{n} \right);$$

$$Z_1(S) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{\nu(X_j)}{n} \right) = \frac{1}{n^2} \sum_{i=1}^{k} n_i^2.$$

Заметим, что при любом τ предложенная мера концентрации имеет минимальное значение, равное $1/n$, при разбиении исследуемого множества на n одноточечных кластеров, и максимальное значение, равное 1, при объединении всех исходных наблюдений в один общий кластер.

При конструировании и сравнении различных кластер-процедур полезно иметь в виду, что объединение двух кластеров S_l и S_m в один дает прирост меры концентрации $Z_1(S)$, равный

$$\Delta Z_1 = \frac{1}{n^2} [(n_l + n_m)^2 - n_l^2 - n_m^2] = \frac{2 n_l n_m}{n^2}$$

Определение *средней меры внутриклассового рассеяния* $I_\tau^{(K)}(S)$ также опирается на понятие степенного среднего. В частности, полагают

$$I_\tau^{(K)}(S) = \left\{ \frac{1}{n} \sum_{i=1}^{k} n_i \left[Q_\tau^{(K)}(S_i) \right]^\tau \right\}^{\frac{1}{\tau}}, \qquad (5.18)$$

где под

$$Q_\tau^{(K)}(S_i) = \left[\frac{1}{n_i^2} \sum_{X_j \in S_i} \sum_{X_l \in S_i} d^\tau(X_j, X_l) \right]^{\frac{1}{\tau}}$$

понимается обобщенная средняя мера рассеяния, характеризующая класс S_i. Числовой параметр τ здесь, как и прежде, выбирается по усмотрению исследователя. Полагая

$$Q_\tau^{(K)}(X) = \left[\frac{1}{\nu(X)} \sum_{X_l \in S(X)} d^\tau(X, X_l) \right]^{\frac{1}{\tau}},$$

где, как и прежде, $S(X)$— кластер, в который входит наблюдение X, а $\nu(X)$ — число элементов в кластере $S(X)$, формулу (5.18) можно переписать в виде

$$I_\tau^{(K)}(S) = M_\tau\left(Q_\tau^{(K)}(X_1),\ \ldots,\ Q_\tau^{(K)}(X_n)\right) =$$

$$= \left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\nu(X_i)}\sum_{X_l \in S(X_i)} d^\tau(X_i,\ X_l)\right]^{\frac{1}{\tau}} \qquad (5.19)$$

При конструировании и сравнении различных кластер-процедур полезно иметь в виду, что объединение двух кластеров S_l и S_m в один дает прирост величины $n\,[I_\tau^{(K)}(S)]^\tau$, непосредственно характеризующей среднюю меру внутриклассового рассеяния, равный

$$\Delta\left[n\,(I_\tau^{(K)})^\tau\right] = \frac{n_l\,n_m}{n_l+n_m}\left\{2\left[\rho_\tau^{(K)}(S_l,\ S_m)\right]^\tau - \right.$$

$$\left. - \left[Q_\tau^{(K)}(S_l)\right]^\tau - \left[Q_\tau^{(K)}(S_m)\right]^\tau\right\}.$$

Очевидно, если ориентироваться на сокращение числа кластеров при наименьших потерях в отношении внутриклассового рассеивания, не обращая внимания на меру концентрации, то естественно объединять два кластера, для которых минимальна величина $\Delta\,(n\,[I_\tau^{(K)}]^\tau)$. Если же одновременно ориентироваться и на рост взвешенной концентрации $Z_1(S)$, то объединение кластеров следует подчинить требованию минимизации величины

$$\frac{\Delta\left[n\left(I_\tau^{(K)}\right)^\tau\right]}{\Delta Z_1(S)} =$$

$$= \frac{\left\{2\left[\rho_\tau^{(K)}(S_l,\ S_m)\right]^\tau - \left[Q_\tau^{(K)}(S_l)\right]^\tau - \left[Q_\tau^{(K)}(S_m)\right]^\tau\right\}^\tau}{n_l+n_m}.$$

5.4.3. **Формулировка экстремальных задач разбиения исходного множества объектов на классы при неизвестном числе классов.** Возможно множество вариантов таких формулировок. Рассмотрим два из них, относящиеся к наиболее естественным и часто используемым.

Вариант 1: *комбинирование функционалов качества.* Требуется найти такое разбиение S^*, для которого некоторая алгебраическая комбинация функционала, характеризующего среднее внутриклассовое рассеяние (5.19), и функционала, характеризующего меру концентрации полученной структуры (5.17), достигала бы своего экстремума. В качестве примеров можно привести комбинации $Q(S)$ и $Q'(S)$,

задаваемые формулами (5.15) и (5.16), а также более общие выражения вида

$$\alpha \cdot I_1(S) + \beta I_2(S) \quad \text{и} \quad [I_1(S)]^{\alpha} \cdot [I_2(S)]^{\beta}, \tag{5.20}$$

где $I_1(S) = I_\tau^{(K)}(S)$; $I_2(S) = \dfrac{1}{Z_\tau(S)}$; α и β — некоторые положительные константы, например $\alpha = \beta = 1$.

Вариант 2: *двойственная формулировка*. Требуется найти разбиение S^*, которое, обладая концентрацией $Z_\tau(S^*)$, не меньшей заданного порогового значения Z_0, давало бы наименьшее внутриклассовое рассеяние $I_\tau^{(K)}(S^*)$, или, в двойственной подстановке: при заданном пороговом значении I_0 найти разбиение S^* с внутриклассовым рассеянием $I_\tau^{(K)}(S^*) \leqslant I_0$ и наибольшей концентрацией $Z_\tau(S^*)$.

5.4.4. Общий вид функционала качества разбиения как функции ряда параметров, характеризующих межклассовую и внутриклассовую структуру наблюдений. Зададимся вопросом: нельзя ли выделить такой достаточно полный набор величин $u_1(S)$, $u_2(S)$, ..., характеризующих как межклассовую, так и внутриклассовую структуру наблюдений при каждом фиксированном разбиении на классы S, чтобы существовала некоторая функция $Q(u_1, u_2, \ldots)$ от этих величин, которую мы могли бы считать в каком-то смысле универсальной характеристикой качества разбиения. В частности, в качестве таких величин $u_1 = u_1(S)$, $u_2 = u_2(S)$,... можно рассмотреть, например, некоторые числовые характеристики: степени близости элементов внутри классов (u_1); степени удаленности классов друг от друга (u_2); степени «одинаковости» распределения многомерных наблюдений внутри классов (u_3); степени равномерности распределения общего числа классифицируемых наблюдений n по классам (u_4).

Что касается установления общего вида функции $Q(u_1, u_2, u_3, u_4)$, то без введения дополнительной априорной информации о наблюдениях X_i (характере и общем виде их закона распределения внутри классов и т п) единственным возможным подходом в решении этой задачи, как нам представляется, является экспертно-экспериментальное исследование Именно с этих позиций в [63] сделана попытка определения общего вида функции Q. Чтобы определить рассмотренные в этой работе величины u_1, u_2, u_3 и u_4, введем понятие *кратчайшего незамкнутого пути* (КНП), соединяющего все n точек исходной совокупности в связный неориентированный граф с минимальной суммарной длиной ре-

бер [1]. Под длиной ребра понимается расстояние между соответствующими точками совокупности в смысле выбранной метрики. Построение такого графа можно начать с пары наиболее близких точек. Если таких пар несколько, то выбирается любая из этих пар. Пусть это будут наблюдения с номерами i_0 и j_0. Затем с помощью сравнения расстояний $d(X_{i_0}, X_j)$ $(j = 1, 2, \ldots, n;\ j \neq i_0,\ j \neq j_0)$ и $d(X_{j_0}, X_q)$, где $q = 1, 2, \ldots, n;\ q \neq j_0$ и $q \neq i_0$, определяются точки $X_{m(i_0)}$ и $X_{m(j_0)}$, наименее удаленные соответственно от точек X_{i_0} и X_{j_0}, и выбирается ближайшая из них X_{m_0}, т. е. $X_{m_0} = X_{m(i_0)}$, если $d(X_{i_0}, X_{m(i_0)}) < d(X_{j_0}, X_{m(j_0)})$, и $X_{m_0} = X_{m(j_0)}$, если $d(X_{j_0}, X_{m(j_0)}) < d(X_{i_0}, X_{m(i_0)})$[2]. Затем точка X_{m_0} «пристраивается» к той из точек X_{i_0} и X_{j_0}, к которой она ближе. Далее сравниваются расстояния $d(X_{i_0}, X_j)$, $d(X_{j_0}, X_q)$ и $d(X_{m_0}, X_\nu)$ $(j, q, \nu \neq i_0;\ j, q, \nu \neq j_0$ и $j, q, \nu \neq m_0)$ и т. д. Очевидно, «разрубая» s ребер такого графа, будем делить совокупность на $s + 1$ классов.

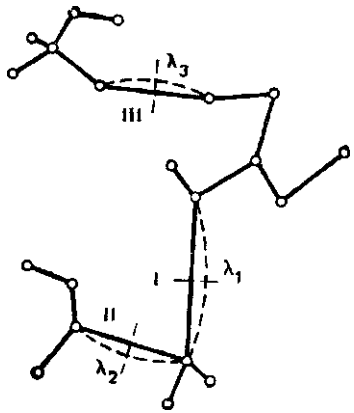


Рис. 5.2. Графическое изображение кратчайшего незамкнутого пути

Пусть $\rho_i(l)$ — i-е ребро части графа, отнесенной к l-му классу. Всего таких ребер, как легко видеть, будет $n_l - 1$. И пусть $\rho^{(l)}_{\min}(\rho)$ — минимальное из ребер, непосредственно примыкающих к ребру ρ и относящихся к l-му классу, если таковое имеется. Пронумеруем в определенном порядке граничные, («разрубленные») ребра $\lambda_1, \lambda_2, \ldots, \lambda_{k-1}$ таким образом, чтобы имелось взаимно однозначное соответствие между номерами граничных ребер и номерами примыкающих к ним классов, за исключением одного, геометрически представленного одним из «хвостов» графа. Выбрасывая ребра I, II, III (рис. 5.2), получаем четыре связных графа, что

[1] Методы классификации, основанные на КНП, используются для решения задач в области антропологии, биологии, сельского хозяйства, лингвистики (см., например: Czekanowski J. Zur Differentialdiagnose der Neandertalgruppe // Kor.-blatt Deutsch. Ges. Antrop. — 1909. — XL. — S. 44—47; Florek K., Lukaszewicz J., Perkal H., Zubzucki S. Sur la liaison et la division des points d'un ensemble fini // Coll. Math. — 1951. — 2. — P. 282—285).

[2] Если $d(X_{i_0}, X_{m(i_0)}) = d(X_{j_0}, X_{m(j_0)})$, то в качестве X_{m_0} можно выбрать любую из точек $X_{m(i_0)}$ и $X_{m(j_0)}$.

соответствует разбиению совокупности на четыре группы. Обозначим с помощью λ_l одно из таких ребер l-го класса.

Теперь, следуя [63], определим величины u_i таким образом:

$$u_1 = \frac{1}{k} \sum_{l=1}^{k} \bar{\rho}(l),$$

где $\bar{\rho}(l) = (\sum_{i=1}^{n_l} \rho_i(l))/(n_l - 1)$ — средняя длина ребер l-го класса;

$$u_2 = \frac{1}{k-1} \sum_{l=1}^{k-1} \lambda_l;$$

$$u_3 = \frac{1}{k-1} \sum_{l=1}^{k-1} \frac{\rho_{\min}^{(l)}(\lambda_l)}{\lambda_l};$$

$$u_4 = k^k \prod_{l=1}^{k} \frac{n_l}{n}.$$

Эмпирический перебор различных вариантов общего вида функции в сочетании с анализом результатов экспертных оценок качества всевозможных разбиений привели авторов [63] к следующей формуле для функционала:

$$Q(S) = \ln \left\{ \frac{[u_2(S)]^a [u_4(S)]^b}{(1 + [u_1(S)]^c)(1 + [u_3(S)]^d)} \right\}, \qquad (5.21)$$

где a, b, c и d — некоторые неотрицательные параметры, оставляющие исследователю определенную свободу выбора в каждом конкретном случае. Авторы [63] отмечали хорошее согласие своих экспериментов с экспертными оценками при $a = b = c = d = 1$.

Из смысла величин u_i $(i = 1, 2, 3, 4)$ следует, что лучшим разбиениям соответствуют большие численные значения функционала Q, так что в данном случае требуется найти такое разбиение S^*, при котором $Q(S^*) = \max\limits_{S} Q(S)$.

Конечно, данный выбор количественного и качественного состава величин u_i и, в еще большей степени, их точное определение являются чисто эвристическими и подчас просто спорными. Это относится в первую очередь к величине u_3. Поэтому читатель должен принимать описанную здесь схему не как рекомендацию к универсальному использованию функционалов типа (5.21) в задачах кластер-анализа, а

лишь как описание конкретного примера одного из возможных подходов при выборе функционалов качества разбиения.

5.4.5. Функционалы качества и необходимые условия оптимальности разбиения. Естественно попытаться проследить, в какой мере выбор того или иного вида функционала качества определяет класс разбиений, в котором следует искать оптимальное. Приведем здесь некоторые результаты, устанавливающие такого рода соответствие.

Будем предполагать используемую метрику евклидовой. Обозначим через $\mathbf{e} = (e_1, \ldots, e_k)$ группу из k p-мерных векторов e_j $(j = 1, \ldots, k)$, а через $S(\mathbf{e}) = (S_1(\mathbf{e}), \ldots, S_k(\mathbf{e}))$ — так называемое *минимальное дистанционное разбиение*, порождаемое точками $\mathbf{e} = (e_1, \ldots, e_k)$, а именно

$$S_1(\mathbf{e}) = \{X : d(X, e_1) \leqslant d(X, e_j),\ j = 2, \ldots, k\};$$

$$S_2(\mathbf{e}) = \overline{S}_1(\mathbf{e}) \cap \{X : d(X, e_2) \leqslant d(X, e_j),\ j \neq 2\};$$

. . . . . .

$$S_k(\mathbf{e}) = \overline{S}_1(\mathbf{e}) \cap \ldots \cap \overline{S}_{k-1}(\mathbf{e}) \cap \{X : d(X, e_k) \leqslant d(X, e_j),\ j \neq k\}.$$

(Здесь и в дальнейшем $\overline{S}_i$ означает дополнение множества S_i до всего пространства $\Pi^p(X)$, т. е. совокупность всех наблюдений (элементов пространства $\Pi^p(X)$, не входящих в состав S_i.) Таким образом, класс $S_j(\mathbf{e})$ состоит из тех точек пространства $\Pi^p(X)$, которые ближе к e_j, чем ко всем остальным e_i $(i \neq j)$. Если для некоторых точек из $\Pi^p(X)$ самыми близкими являются сразу несколько векторов e_j $(j = 1, \ldots, k)$, то эти точки относим к классу с минимальным индексом.

Разбиение $S = (S_1, \ldots, S_k)$ называется *несмещенным*, если это разбиение с точностью до множеств меры нуль совпадает с минимальным дистанционным разбиением, порождаемым векторами средних $\overline{X}_l = \frac{1}{P_l} \int\limits_{S_l} XP(\mathrm{d}X)$, где $P_l = \int\limits_{S_l} P(\mathrm{d}X)$.

Утверждение 1 (для функционалов типа Q_1 (см. (5.11))). Минимальное значение функционала $Q_1(S) = \sum\limits_{l=1}^{k} \int\limits_{S_l} \rho^2(X, \overline{X}_l) P(\mathrm{d}X)$ достигается только на несмещенных разбиениях. Это означает, что оптимальное разбиение обязательно должно быть несмещенным [166].

Следующий результат относится к довольно широкому классу функционалов качества разбиения совокупности

на два класса. Разбиение на два класса может быть задано с помощью так называемой разделяющей функции. А именно точки пространства $\Pi^p(X)$, на которых разделяющая функция принимает неотрицательное значение, относятся к одному классу, а остальные — к другому. Поэтому поиск класса оптимальных разбиений в этом случае эквивалентен поиску класса оптимальных разделяющих функций.

Для иллюстрации дальнейшего изложения будем рассматривать вероятностную модификацию функционала Q_2' (см. (5.14)).

Пусть расстояние $d(X, Y)$ задается с помощью соотношения (5.3) потенциальной функцией вида

$$K(X, Y) = \sum_{i=1}^{N} \lambda_i^2 \varphi_i(X) \varphi_i(Y),$$

где $\varphi_i(X)$ $(i = 1, \ldots, N)$ — некоторая система функций на $\Pi^p(X)$.

Функционал Q_2' через потенциальную функцию $K(X,Y)$ выражается следующим образом:

$$Q_2'(S) = 2 \int_{\Pi^p(X)} K(X, X) P(\mathrm{d}X) - 2 \frac{1}{P_1} \int_{S_1} \int_{S_1} K(X, Y) \times$$

$$\times P(\mathrm{d}X) P(\mathrm{d}Y) - 2 \frac{1}{P_2} \int_{S_2} \int_{S_2} K(X, Y) P(\mathrm{d}X) P(\mathrm{d}Y).$$

Поскольку в правой части этого равенства первый интеграл не зависит от разбиения, то минимум функционала $Q_2'(S)$ достигается на тех разбиениях, на которых функционал

$$\overline{Q}_2(S) = \frac{1}{P_1} \int_{S_1} \int_{S_1} K(X, Y) P(\mathrm{d}X) P(\mathrm{d}Y) +$$

$$+ \frac{1}{P_2} \int_{S_2} \int_{S_2} K(X, Y) P(\mathrm{d}X) P(\mathrm{d}Y)$$

достигает максимума.

Введем в рассмотрение *спрямляющее пространство* $\Pi^N(Z)$, координаты $z^{(i)}$ векторов $Z \in \Pi^N(Z)$ которого определяются соотношениями

$$z^{(i)} = \lambda_i \varphi_i(X) \quad (i = 1, \ldots, N).$$

В спрямляющем пространстве $\Pi^N(Z)$ вероятностной мерой P, заданной в исходном пространстве $\Pi^p(X)$, индуцируется

своя вероятностная мера $P^{(Z)}$. Однако в целях упрощения обозначений будем опускать верхний индекс Z у этой новой меры. Функционал $Q_2(S)$ в спрямляющем пространстве примет вид

$$\overline{Q}_2(S) = \frac{1}{P_1}\left[\int_{S_1} ZP(\mathrm{d}Z)\right]^2 + \frac{1}{P_2}\left[\int_{S_2} ZP(\mathrm{d}Z)\right]^2.$$

Пусть

$$M_i^{(\nu)} = \int_{S_i} Z^\nu P(\mathrm{d}Z) \quad (\nu = 0,\ 1,\ \ldots,\ r;\quad i = 1,\ 2).$$

Здесь $Z^{2j} = [(Z, Z)]^j$ — числа, $Z^{2j+1} = [(Z, Z)]^j Z$ — векторы.

В работе [32] формулируется утверждение, устанавливающее класс функций в спрямляющем пространстве $\Pi^N(Z)$, среди которых следует искать разделяющую функцию, доставляющую экстремум функционалу качества разбиения. Показано, что если функционал качества Φ является дифференцируемой функцией от M_i^ν ($\nu = 1, \ldots, r$), а вероятностное распределение P^Z сосредоточено на ограниченном множестве и обладает непрерывной плотностью, то в случае достижения функционалом Φ своего экстремума на некоторой разделяющей функции, этот же экстремум достигается на разделяющей функции, являющейся полиномом r-й степени вида:

$$f(Z) = \sum_{\nu=1}^{r} (c_\nu,\ Z^\nu),$$

где

$$c_\nu = \frac{\partial \Phi}{\partial M_1^{(\nu)}} - \frac{\partial \Phi}{\partial M_2^{(\nu)}},$$

а (c_ν, Z^ν) означает при четном ν произведение чисел c_ν и Z^ν, а при нечетном ν — скалярное произведение векторов c_ν и Z^ν

Утверждение 2 (для функционалов типа $Q_2'(S)$ и разбиений на два класса). Для функционала Q_2' утверждение 1 означает, что класс разделяющих функций, среди которых надо искать наилучшее в спрямляющем пространстве разбиение, имеет вид

$$f(Z) = (c,\ Z) - a,$$

где

$$c = \frac{\partial Q_2'}{\partial M_1} - \frac{\partial Q_2'}{\partial M_2} = 2\left(\frac{M_1}{P_1} - \frac{M_2}{P_2}\right); \qquad (5.22)$$

$$a = \frac{\partial Q_2'}{\partial P_1} - \frac{\partial Q_2'}{\partial P_2} = \left(\frac{M_1}{P_1}\right)^2 - \left(\frac{M_2}{P_2}\right)^2;$$

$M_i = M_i^{(1)}$ $(i = 1, 2)$.

Класс разделяющих функций в спрямляющем пространстве очевидным образом определяет класс разделяющих функций в исходном пространстве $\Pi^p(X)$.

Если $K(X, Y) - (X, Y)$ является скалярным произведением векторов X и Y, то спрямляющее пространство $\Pi^N(Z)$ совпадает с исходным пространством $\Pi^p(X)$, а метрика, задаваемая потенциальной функцией $K(X, Y)$, совпадает с обычной евклидовой метрикой. Функционалы Q_2 и Q_2', рассматриваемые относительно этой метрики, совпадают с точностью до константы. В этом случае, как нетрудно видеть, разбиение, задаваемое разделяющей функцией $f(Z)$, является несмещенным разбиением.

5.4.6. Функционалы качества разбиения как результат применения метода максимального правдоподобия к задаче статистического оценивания неизвестных параметров. Приведем пример, иллюстрирующий возможность обоснования выбора общего вида функционала качества разбиения на классы в ситуациях, в которых исследователю удается описать механизм генерирования анализируемых данных одной из классических моделей.

Пусть априорные сведения позволяют определить i-й однородный класс (кластер) как нормальную генеральную совокупность наблюдений с вектором средних $\mathbf{a}_i$ и ковариационной матрицей Σ_i. При этом $\mathbf{a}_i$ и Σ_i, вообще говоря, неизвестны. Известно лишь, что каждое из наблюдений X_1, X_2, ..., X_n извлекается из одной из k нормальных генеральных совокупностей $N(\mathbf{a}_i, \Sigma_i)$ (i 1,2, ..., k). Задача исследователя — определить, какие n_1 из n исходных наблюдений извлечены из класса $N(\mathbf{a}_1, \Sigma_1)$, какие n_2 наблюдений извлечены из класса $N(\mathbf{a}_2, \Sigma_2)$ и т. д. Очевидно, числа $n_1, n_2, \ldots, n_k$, вообще говоря, также неизвестны.

Если ввести в рассмотрение вспомогательный вектор «параметр разбиения» $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$, в котором компонента определяет номер класса, к которому относится наблюдение X_i, т. е. $\gamma_i = l$, если $X_i \in N(\mathbf{a}_l, \Sigma_l)$ $(i = 1,2, \ldots, n)$, то задачу разбиения на классы можно формулировать как задачу оценивания неизвестных параметров $\gamma_1, \gamma_2, \ldots, \gamma_n$ при «мешающих» неизвестных параметрах $\mathbf{a}_l$ и Σ_l

(l = 1,2,..., k). Обозначив весь набор неизвестных параметров с помощью θ, т. е. $\theta = (\gamma; \mathbf{a}_1, \ldots, \mathbf{a}_k; \Sigma_1, \ldots, \Sigma_k)$, и пользуясь известной [16] техникой, получаем логарифмическую функцию правдоподобия для наблюдений X_1, X_2, ..., X_n:

$$l(\theta) = -\frac{1}{2}\sum_{l=1}^{k}\left[\sum_{X_i \in S_l(\gamma)}(X_i - \mathbf{a}_l)'\,\Sigma_l^{-1}(X_i - \mathbf{a}_l) + \right.$$

$$\left. + n_l(\gamma)\log|\Sigma_l|\right]. \qquad (5.23)$$

Как известно, оценка $\widehat{\theta}$ параметра θ по методу максимального правдоподобия находится из условия $l(\widehat{\theta}) = \max l(\theta)$.

Поэтому естественно было бы попытаться найти такое значение «параметра разбиения» γ, а также такие векторы средних $\widehat{\mathbf{a}}_l$ и ковариационные матрицы $\widehat{\Sigma}_l$, при которых величина $-2l(\theta)$ достигла бы своего абсолютного минимума[1].

При известном разбиении γ оценками максимального правдоподобия для $\mathbf{a}_l$ будут «центры тяжести» классов

$$\overline{X}^{(\gamma)}(l) = \frac{1}{n_l}\sum_{X_i \in S_l} X_i \quad (l = 1, 2, \ldots, k).$$

Подставляя их в (5.23) вместо $\mathbf{a}_l$ и воспользовавшись очевидными тождественными преобразованиями, приходим к эквивалентности задачи поиска минимума функции $-2l(\theta)$, определенной соотношением (5.23), и задачи поиска минимума выражения

$$\sum_{l=1}^{k}\left[\sum_{X_i \in S_l}(X_i - \overline{X}^{(\gamma)}(l))'\,\Sigma_l^{-1}(X_i - \overline{X}^{(\gamma)}(l)) + n_l\log|\Sigma_l|\right], \qquad (5.24)$$

или, что то же, выражения

$$\sum_{l=1}^{k}|\mathrm{Sp}(n_l W_l \Sigma_l^{-1}) + n_l l\log|\Sigma_l||. \qquad (5.25)$$

В последнем выражении W_l — выборочная ковариационная матрица, вычисленная по наблюдениям, входящим в состав l-го класса (см. (5.13)).

[1] Оговоримся сразу, что даже факт состоятельности полученных при этом оценок $\widehat{\theta}$ остается под сомнением, поскольку размерность неизвестного векторного параметра θ превосходит в данном случае общее число наблюдений, которыми располагаем.

Анализ выражений (5.24) и (5.25) в некоторых частных случаях немедленно приводит к следующим интересным выводам:

если ковариационные матрицы исследуемых генеральных совокупностей равны между собой и известны, то задача оценивания неизвестного параметра θ по методу максимального правдоподобия равносильна задаче разбиения наблюдений X_i на классы, подчиненной функционалу качества разбиения вида $Q_1(S)$, в котором под расстоянием d подразумевается расстояние Махаланобиса;

если ковариационные матрицы исследуемых генеральных совокупностей равны между собой, но неизвестны, то, подставляя в (5.25) вместо $\mathbf{\Sigma}_l = \mathbf{\Sigma}$ ее оценку максимального правдоподобия

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{l=1}^{k} n_l \mathbf{W}_l,$$

убеждаемся в эквивалентности задачи оценивания (по методу максимального правдоподобия) параметра θ и задачи поиска разбиения наблюдений X_i на классы, наилучшего в смысле функционала качества $Q_3(S)$;

если ковариационные матрицы исследуемых генеральных совокупностей не равны между собой и не известны, то, подставляя в (5.25) вместо $\mathbf{\Sigma}_l$ их оценки максимального правдоподобия $\mathbf{W}_l$, убеждаемся в эквивалентности задачи оценивания по методу максимального правдоподобия параметра θ и задачи поиска разбиения наблюдений X_i на классы, наилучшего в смысле функционала качества $Q_1(S)$.

В [303] авторы пытаются конструировать алгоритмы, реализующие идею получения оценок максимального правдоподобия для параметра θ. Однако, нам представляется, главная ценность подобного подхода лишь в его методологической, качественной стороне, в том, что он позволяет строго осмыслить и формализовать некоторые функционалы качества разбиения, введенные ранее чисто эвристически. Конструктивная же сторона подобного подхода упирается в трудно преодолимые препятствия вычислительного плана, связанные с колоссальным количеством переборов вариантов уже при сравнительно небольших размерностях p и объемах выборки.

Роль функционалов качества разбиения в построении общей теории автоматической классификации и разнообразные примеры функционалов качества, используемых в конкретных (распространенных в статистической практике) алгоритмах автоматической классификации, описаны в гл. 10.

5.4.7. Функционалы качества классификации как показатели степени аппроксимации данных. Поскольку формируемая классификационная структура является приближенным представлением имеющихся данных, естественно возникает идея формализации задачи классификации как задачи аппроксимации матрицы данных матрицей, характеризующей искомую классификацию. Для реализации этой идеи необходимо тем или иным образом выбрать, во-первых, матричный способ представления классификации и, во-вторых, меру близости между матрицами, соответствующими исходным данным и искомым классификациям. Рассмотрим некоторые возникающие при этом задачи для ситуаций, когда классификация рассматривается как а) иерархия классов, б) разбиение, в) совокупность непустых (возможно, пересекающихся) подмножеств, а в качестве меры близости матриц используется обычная евклидова метрика — сумма квадратов разностей их соответствующих элементов.

В одной из первых версий аппроксимационная постановка задачи была рассмотрена для классификации, представленной иерархической системой кластеров, т. е. совокупностью разбиений S^1, S^2,..., S^m множества объектов O, такой, что каждое последующее разбиение S^t получается из предыдущего разбиения S^{t-1} объединением некоторых его классов ($t = 1,2, \ldots, m$), причем последнее разбиение S^m состоит из единственного класса, совпадающего со всем множеством O[1]. Разбиение S^t соответствует t-му уровню иерархии, характеризуемому также *индексом уровня* u_t, монотонно зависящим от t (обычно полагают $u_t = t$). Такая иерархическая система порождает, в частности, матрицу расстояний между объектами O_i, O_j по следующему правилу: расстояние $d(i, j)$ между ними равно индексу уровня u_t того разбиения S^t, в котором они впервые попадают в один и тот же класс. Мера близости $d(i, j)$ является *ультраметрикой*, т. е. удовлетворяет усиленному неравенству треугольника: $d(i, j) \leqslant \max(d(i, k), d(j, k))$ для любых $i, j, k = 1, \ldots, n$. Имеет место и обратное утверждение: всякая ультраметрика порождает иерархическую систему кластеров, индексами уровня которой служат различающиеся значения ультраметрики. Таким образом, иерархические системы кластеров эквивалентно представимы ультраметрическими матрицами близости. Это позволяет формулировать задачу построения иерархической системы кластеров как задачу аппроксимации заданной матрицы расстояний p_{ij} с помощью ультраметрических матриц [250]. Ока-

[1] Иерархическим процедурам классификации посвящена гл. 8.

залось, что если искомая ультраметрика удовлетворяет априорному ограничению $d(i, j) \leqslant p_{ij}$, а критерий аппроксимации — сумма квадратов (или других монотонных функций) от разностей $d(i, j) - p_{ij}$, то оптимальной будет система кластеров, получаемая по методу «ближайшего соседа».

Иными словами, ультраметрическое расстояние $d(i,j)$ есть не что иное, как минимальное значение p, при котором p-граф $\Gamma_p = \{(i, j): p_{ij} \leqslant p\}$ содержит последовательность ребер, соединяющую объекты O_i и O_j (или, что то же самое, максимальное ребро на той части кратчайшего незамкнутого пути, которая соединяет эти объекты).

Из сказанного вытекает следующее свойство «компактности» оптимальных кластеров: любое «внутреннее» расстояние p_{ij} для объектов O_i, O_j принадлежащих одному и тому же кластеру S, меньше любого расстояния $d(k, l)$ «во вне», т. е. при любых O_k из S и O_l, не принадлежащих S.

Задача аппроксимации с противоположным ограничением $d(i, j) \geqslant p_{ij}$ приводит к методу «дальнего соседа» с аналогичными свойствами.

Задача аппроксимации матрицы близости с помощью иерархической системы кластеров без ограничений не рассматривалась практически в литературе, в отличие от задач аппроксимации с помощью разбиений и подмножеств (задаваемых, впрочем, специальными типами ультраметрик). Рассмотрим некоторые результаты, полученные в [92] применительно к ситуации, когда элементы матрицы характеризуют *не расстояние, а сходство* (связь) между объектами.

Разбиение $S = \{S_1, \ldots, S_k\}$ множества объектов O представим матрицей связи вида $\{\lambda s_{ij}\}$, где $s_{ij} = 1$, если O_i, O_j содержатся в одном и том же классе разбиения S, и $s_{ij} = 0$ в противном случае, а λ — вещественное число, характеризующее уровень связи между объектами «внутри» классов S. Нетрудно видеть, минимизация суммы квадратов разностей величин $p_{ij} - \lambda s_{ij}$ при заданном $\lambda > 0$ эквивалентна задаче отыскания такого разбиения S, которое максимизирует суммарные внутренние связи

$$f(S, \pi) = \sum_{t=1}^{m} \sum_{i,j \in S_t} (p_{ij} - \pi) = \sum_{t=1}^{m} \sum_{i,j \in S_t} p_{ij} - \pi \sum_{t} n_t^2. \quad (5.26)$$

Величина $\pi = \lambda/2$ играет здесь, с одной стороны, роль порога существенности связей p_{ij}, поскольку при $p_{ij} > \pi$ выгодно поместить O_i и O_j в один и тот же класс ($p_{ij} - \pi > 0$), а при $p_{ij} < \pi$, напротив, это невыгодно, и, с другой стороны, роль параметра компромисса двух противоречивых

критериев — максимума суммы внутренних связей $\sum_t \sum_{i,j \in S_t} p_{ij} = I(S)$ и минимума меры концентрации $\frac{1}{n^2} \sum_t n_t^2 = Z_1(S)$.

Оптимальное по данному критерию разбиение $S = \{S_1, \ldots, S_k\}$ компактно в том смысле, что средняя связь $p(S_t) = \sum_{i,j \in S_t} p_{ij}/n_t^2$ в каждом из его классов S_t не меньше, чем средняя связь $p(S_t, \bar{S}_t)$ во вне и чем средняя связь между любыми классами $p(S_l, S_t)$. Точнее, для любых t, q, l справедливо неравенство $p(S_l, \bar{S}_l), p(S_q, S_t) \leqslant \pi \leqslant p(S_l)$.

Разбиение, удовлетворяющее данному свойству, может быть найдено с помощью локально-оптимального агломеративного алгоритма, на каждом шаге которого объединяются те классы S_t, S_q, для которых сумма связей $\sum_{i \in S_t} \sum_{j \in S_q} (p_{ij} - \pi)$ максимальна и положительна. Процесс объединений прекращается, когда такие суммы для всех $t \neq q$ становятся неположительными. Таким образом, данная задача приводит к сумме всех связей (за вычетом порога) как мере близости классов.

В том случае, когда величина λ не фиксирована заранее, а подбирается в соответствии с квадратичным критерием аппроксимации, ее оптимальное значение (при данном S) равно среднему значению внутренних связей

$$\lambda(S) = \sum_{t=1}^{k} \sum_{i,j \in S_t} p_{ij} \Big/ \sum_{t=1}^{k} n_t^2.$$

В этой ситуации необходимое условие оптимальности S может быть уточнено следующим образом: средняя внутренняя связь $\lambda(S)$ по крайней мере вдвое превышает средние связи между любыми двумя классами ($p(S_t, S_q)$) или во вне ($p(S_t, \bar{S}_t)$). Данное условие также может использоваться в качестве критерия оптимизации и остановки в агломеративном методе последовательного объединения классов.

К достоинствам рассмотренной аппроксимационной задачи классификации относится то, что, во-первых, она допускает интерпретацию в терминах порога существенности (он же показатель компромисса) π; во-вторых, приводит к «компактным» в указанном смысле кластерам, число которых не задается заранее, и, в-третьих, при некоторых конкретных способах расчета связей p_{ij} эквивалентна другим аппроксимационным критериям многомерной классификации по матрице «объект — свойство», содержащей как количественные, так и номинальные признаки [111]. Особенно-

стью данного критерия является наличие единого порога существенности для всех классов, что неудобно в тех нередких ситуациях, когда структура данных отражает наличие кластеров различных «диаметров».

Этот недостаток в значительной мере преодолевается в методике последовательной аппроксимации матрицы связи с помощью отдельных подмножеств множества объектов O.

Подмножество $S \subseteq O$ характеризуется $(n \times n)$-матрицей $\{\lambda s_{ij}\}$, где $s_{ij} = 1$ при O_i, $O_j \in S$ и $s_{ij} = 0$ в противном случае. При фиксированном значении $\lambda > 0$ задача квадратичной аппроксимации матрицы $\{p_{ij}\}$ в множестве матриц такого вида эквивалентна задаче максимизации суммы внутренних связей

$$f(S, \pi) = \sum_{i,j \in S} (p_{ij} - \pi) = \sum_{i,j \in S} p_{ij} - \pi n_S,$$

что представляет собой часть критерия (5.26), соответствующую одному классу. Здесь $\pi = \lambda/2$, а n_S — число объектов в S.

Необходимое условие оптимальности здесь принимает следующий вид. Оптимальное множество S является π-кластером в том смысле, что для всякого объекта $O_i \in S$ его средняя связь $p(i, S) = \sum_{j \in S} p_{ij}/n_S$ с S не меньше, чем π, тогда как для $O_i \notin S$ $p(i, S) \leqslant \pi$. Это свойство позволяет формировать локально-оптимальный π-кластер путем последовательного присоединения к S, начиная с $S = \varnothing$, объектов, имеющих с «текущим» S максимальную среднюю связь $p(i, S)$, если $p(i, S) > \pi$.

В том случае, когда величина λ или $\pi = \lambda/2$ заранее не задана, ее выбор можно также подчинить аппроксимационному критерию, который, как нетрудно убедиться, равен в этом случае $g^2(S)$, где

$$g(S) = \sum_{i,j \in S} p_{ij}/n_S = n_S p(S), \tag{5.27}$$

причем $p(S) = \sum_{i,j \in S} p_{ij}/n_S^2$ — не что иное, как оптимальное значение λ, равное средней внутренней связи.

Оптимальное по данному критерию множество S является «сильным» кластером в том смысле, что его средняя внутренняя связь $p(S)$ по крайней мере вдвое превышает среднюю связь $p(i, S)$ с S любого объекта $O_i \notin S$. Отсюда вытекает алгоритм построения локально-оптимального сильного кластера S с помощью последовательных присоединений к «текущему» S тех O_i, для которых $p(i, S)$ максимально (по $i \notin S$), если $p(i, S) > p(S)/2$. Подобные алгоритмы формирования отдельных кластеров развиваются в литера-

туре довольно давно, правда, из чисто эвристических соображений, например метод β-коэффициентов [161], алгоритм Спектр [34] и др.

В случае, когда часть связей p_{ij} может быть отрицательна (если, например, матрица $\{p_{ij}\}$ предварительно центрирована), аппроксимационный критерий $g^2(S)$ допускает решение, при котором $g(S)$, а значит, и порог λ, отрицательны. В этом случае оптимальное множество S является «антикластером», так как состоит из наиболее непохожих друг на друга объектов. До последнего времени задачи построения антикластеров в приложениях рассматривались редко.

Описанный подход допускает естественное развитие, состоящее в многократном повторении аппроксимации применительно к «остаточным» матрицам связей, получаемым вычитанием из p_{ij} «объясненных» на данном шаге связей λs_{ij}. Очевидно, на разных этапах процесса аппроксимируемые матрицы связи будут разными, так что конструируемые кластеры и их средние внутренние связи будут, вообще говоря, разными. Указанный метод последовательного исчерпания связей p_{ij} (качественный факторный анализ [110]) может рассматриваться как пошаговая процедура идентификации модели аддитивных кластеров, согласно которой заданная матрица связи $\{p_{ij}\}$ представляется (с некоторой погрешностью) в виде суммы матриц $\{\lambda_t s_{ij}^t\}$, отвечающих искомым кластерам S_t:

$$p_{ij} = \lambda_0 + \lambda_1 s_{ij}^1 + \lambda_2 s_{ij}^2 + \ldots + \lambda_h s_{ij}^k + \varepsilon_{ij}, \qquad (5.28)$$

где ε_{ij} — минимизируемые невязки модели.

Согласно методу последовательного исчерпания сначала определяется и вычитается из p_{ij} оптимальное λ_0, равное среднему значению p_{ij}, затем матрица $p_{ij} - \lambda_0$ аппроксимируется матрицей вида $\{\lambda_1 s_{ij}^1\}$; после этого остаточная матрица $p_{ij} - \lambda_0 - \lambda_1 s^1{}_{ij}$ аппроксимируется матрицей $\{\lambda_2 s_{ij}{}^2\}$ и т. д. Особенностью метода является декомпозиция

$$\sum_{i,j} p_{ij}^2 = \sum_t \nu_t^2 + \sum_{i,j} \varepsilon_{ij}^2, \qquad (5.29)$$

где

$$\nu_t = g(S_t) = n_t p(S_t),$$

справедливая независимо от того, пересекаются кластеры S_t друг с другом или нет.

Разложение (5.29) позволяет трактовать величины v_t как характеристики значимости вклада отдельных кластеров в дисперсию исходных связей и на этой основе оценивать доли объясненной и необъясненной дисперсии данных, что включает кластер-анализ в традиционную методологию многомерного статистического анализа. Указанные свойства метода последовательного исчерпания (декомпозиция (5.29), «сильная компактность» кластеров, вычислительная эффективность) дают ему определенные преимущества по сравнению с другими подходами к оценке параметров модели (5.28) [111].

Рассмотрим некоторые критерии классификации, возникающие в задачах аппроксимации таблиц «объект — свойство». В этих задачах подмножества объектов S задаются n-мерными (1, 0)-векторами $z = (z_i)$, где $z_i = 1$, если $x_i \in S$ и $z_i = 0$ в противном случае. Разбиение $S = \{S_1, \ldots, S_k\}$ множества объектов задается $(n \times k)$-матрицей $Z = (z_{it})$ со столбцами z_1, z_2, z_k, характеризующими классы разбиения S. Очевидно, $(n \times n)$-матрица $\mathbf{ZZ}'$ совпадает с ранее определенной матрицей $\{s_{ij}\}$, а $k \times k$-матрица $\mathbf{Z}'\mathbf{Z}$ — диагональная матрица величин n_t. Линейное пространство $L(Z) = \{z/z = Za$ при подходящем $a\}$ является адекватным представлением разбиения S, поскольку всякий вектор $z = Za$ выражает возможную кодировку классов S_t величинами a_t.

Аппроксимационные построения в терминах линейных пространств, ассоциированных с номинальными и количественными признаками, приводят к критериям вида (5.1), (5.11) и (5.26) при подходящих способах вычисления характеристик близости объектов [111]. Здесь рассмотрим только наиболее прозрачную схему *метода главных кластеров* [112]. Согласно этой схеме матрица данных $\mathbf{X} = (x_{ij})$ (i — номера объектов, j — номера признаков) аппроксимируется каким-либо элементом $z \in L(Z)$, где Z — $(n \times k)$-матрица искомой классификации (с не обязательно непересекающимися классами), так что справедливо представление

$$x_{ij} = a_{1j} z_{i1} + a_{2j} z_{i2} + \ldots + a_{kj} z_{ik} + \varepsilon_{ij} \qquad (5.30)$$

с «невязками» ε_{ij}.

Смысл равенств (5.30): значения признаков j на объектах $O_i \in S_t$ с точностью до невязок ε_{ij} совпадают с величинами a_{tj}, которые задают, таким образом, «эталон» кластера S_t в признаковом пространстве.

Модель (5.30) является *линейной моделью факторного анализа* (см. гл. 14) с той особенностью, что величины z_{it} принимают не любые значения, а только нулеединичные.

Аппроксимационная задача отыскания произвольных a_{tj} и нулеединичных z_{it} по заданным x_{ij} в модели (5.30) с целью минимизации суммы квадратов невязок

$$\sum_{i,j} \varepsilon_{ij}^2, \tag{5.31}$$

как нетрудно показать, эквивалентна задаче построения классификации $S = \{S_1, \ldots, S_k\}$ (при непересекающихся классах), минимизирующей критерий (5.11) с $a_{tj} = \bar{x}_j(t)$, или, что то же самое, максимизирующей критерии вида

$$g(S) = \sum_{t=1}^{k} \frac{1}{n_t} \sum_{i,j \in S} b_{ij}, \tag{5.32}$$

где связи между объектами — их скалярные произведения $b_{ij} = \sum_t x_{it} x_{jt}$.

Вид критерия (5.31) позволяет в какой-то мере уточнить характер предварительного преобразования данных, необходимого для адекватного применения критериев, эквивалентных (5.11) и (5.32). Действительно, в сумму (5.32) все невязки ε_{ij} входят равноправно, что подразумевает равноправие всех элементов матрицы данных. Скажем, изменение масштаба j-го признака в α раз вызовет α^2-кратное изменение «веса» соответствующих невязок в критерии (5.31) и соответствующее изменение решения. По-видимому, критерию (5.31) соответствует предварительное уравновешивание матрицы данных, приводящее к выполнению равенств $\sum_i x^2_{ij} = n$, $\sum_j x^2_{ij} = p$.

Формулировка задачи классификации как задачи оценки параметров модели (5.30) позволяет распространить на нее принцип пошаговой аппроксимации. На первом шаге отыскиваются параметры первого кластера путем минимизации критерия $\sum_{i,j} (x_{ij} - a_j z_i)^2$ по произвольным a_j и нулеединичным z_i. Полученное решение определяет эталонный вектор первого кластера $a_1 = (a_{1j})$ и его состав $z_1 = (z_{i1})$. На общем $(t+1)$-м шаге исходя из полученного решения a_t, z_t и «текущей» остаточной матрицы x^t_{ij} осуществляется переход к матрице следующего шага $x^{t+1}_{ij} = x^t_{ij} - a_{tj} z_{it}$.

Метод назван методом главных кластеров из-за аналогии с методом главных компонент, «простирающейся» вплоть до декомпозиции

$$\sum_{i,j} x^2_{ij} = \sum_{t=1}^{k} \nu_t + \sum_{i,j} \varepsilon^2_{ij}, \tag{5.33}$$

где $v_t = g(S_t) = n_t\, b(S)$ — вклад t-го кластера S_t в суммарный разброс данных, справедливой даже при пересекающихся главных кластерах. Ситуация здесь похожа на ту, которая обсуждалась применительно к модели аддитивных кластеров (5.28) с декомпозицией (5.29). На каждом t-м шаге аппроксимация сводится к максимизации величины $g(S_t)$ и соответственно каждый главный кластер является сильным кластером (относительно матрицы связей $\{b_{ij}\}$).

Локально-оптимальный алгоритм последовательного присоединения объектов к кластеру S начиная с $S = \emptyset$, по максимуму средней связи $b(i, S)$ $(O_i \notin S)$, имеет здесь простой геометрический смысл: к S присоединяется тот объект O_i, длина проекции которого на радиус-вектор центра тяжести S превышает половину длины радиус-вектора.

С точки зрения интерпретации главных кластеров и отбора «значимых» компонент решения особый интерес представляют разложения

$$v_t = g(S_t) = \sum_{x_i \in S} b(i, S_t) = \sum_j a_{tj}^2 n_t,$$

которые характеризуют относительный вклад (значимость) отдельных объектов $(b(i, S_t))$ и признаков $(a_{tj}^2 n_t)$ применительно к данному конкретному кластеру.

Таким образом, обращение к модели (5.30) расширяет возможности традиционного дисперсионного критерия (5.11) и дает решения, обладающие определенными преимуществами («шлейф» интерпретирующих характеристик, «сильная компактность» кластеров, возможность пересечений и т.п.).

ВЫВОДЫ

1. *Общая постановка задачи классификации* совокупности объектов O_1, O_2, ..., O_n в условиях отсутствия обучающих выборок состоит в требовании разбиения этой совокупности на некоторое число (заранее известное или нет) однородных в определенном смысле классов. При этом исходная информация о классифицируемых объектах представлена либо значениями многомерного признака (по каждому объекту в отдельности), либо матрицей попарных расстояний (или близостей) между объектами, а понятие однородности основано на предположении, что геометрическая близость двух или нескольких объектов означает близость их «физических» состояний, их сходство.

2. *Математическая постановка задачи автоматической классификации* требует формализации понятия «качество разбиения». С этой целью в рассмотрение вводится понятие критерия (функционала) качества разбиения $Q(S)$, который задает способ сопоставления с каждым возможным разбиением S заданного множества объектов на классы некоторого числа $Q(S)$, оценивающего (в определенной шкале) степень оптимальности данного разбиения. Тогда задача поиска наилучшего разбиения S^* сводится к решению оптимизационной задачи вида

$$Q(S) \to \underset{S \in \mathbf{A}}{\text{extr}}, \tag{5.34}$$

где $\mathbf{A}$ — множество всех допустимых разбиений.

3. В зависимости от наличия и характера априорных сведений о природе искомых классов и от конечных прикладных целей исследователь обращается к одной из трех основных составных частей математического аппарата классификации в условиях отсутствия обучающих выборок: 1) *методам расщепления смесей* вероятностных распределений (каждый класс интерпретируется как параметрически заданная одномодальная генеральная совокупность при неизвестном значении определяющего ее параметра, а классифицируемые наблюдения — как выборка из смеси таких генеральных совокупностей), 2) *методам собственно автоматической классификации* или *кластер-анализу* (исследователь не имеет оснований для параметризации модели, а иногда и для интерпретации последовательности классифицируемых наблюдений в качестве выборки из генеральной совокупности); 3) классификационным *процедурам иерархического типа* (главная цель — получение наглядного представления о стратификационной структуре всей классифицируемой совокупности, например в виде дендрограммы).

4. *Выбор метрики* (*или меры близости*) *между объектами*, каждый из которых представлен значениями характеризующего его многомерного признака, является *узловым моментом исследования*, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при любом используемом для этого алгоритме разбиения. В каждой конкретной задаче этот выбор должен производиться по-своему, в зависимости от главных целей исследования, физической и статистической природы анализируемого многомерного признака, априорных сведений о его вероятностной природе и т.п. В этом смысле схемы, основанные на анализе смесей распределений, а также класси-

фикация по исходным данным, уже представленным в виде матрицы попарных расстояний (близостей), находятся в выгодном положении, поскольку не требуют решения вопроса о выборе метрики.

5 Важное место в построении классификационных процедур, в первую очередь иерархических, занимает *проблема выбора способа вычисления расстояния между подмножествами объектов* Изящное обобщение большинства используемых в статистической практике вариантов вычисления расстояний между двумя группами объектов дает расстояние, подсчитываемое как обобщенное степенное среднее всевозможных попарных расстояний между представителями рассматриваемых двух групп (см. (5.8)).

6. В статистической практике выбор функционала качества разбиения $Q(S)$ обычно осуществляется весьма произвольно, опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо точную формализованную схему (см., например, способ вывода функционала качества (5.21)). Однако *ряд распространенных в статистической практике функционалов качества удается постфактум обосновать и осмыслить в рамках строгих математических моделей*. Возможность этого появляется при наличии дополнительных априорных сведений о классах, позволяющих, например, представлять каждый класс в качестве параметрически заданной одномодальной генеральной совокупности (см. основанный на смеси нормальных совокупностей математико-статистический анализ функционалов $Q_1(S)$, $Q_3(S)$ и $Q_4(S)$ в п. 5.4.6)

7. Еще один подход к осмыслению и обоснованию методов автоматической классификации представлен аппроксимационными моделями, когда искомая классификация характеризуется матрицей определенной структуры (например, ультраметрической матрицей близости или аддитивными кластерами (5 28)), а задача состоит в том, чтобы оценить параметры этой структуры таким образом, чтобы она минимально отличалась от матрицы исходных данных. В такой постановке проблема классификации сближается с проблемами факторного анализа (см., в частности, модель главных кластеров (5.30), которая является реализацией для данного случая соотношений линейной модели факторного анализа – - см. гл. 14). Поэтому данный подход в определенной мере интегрирует традиционные методы кластер-анализа («компактность» кластеров в признаковом пространстве) и многомерной статистики (декомпозиция разброса исходных данных на «вклады» отдельных кластеров и других элементов решения).

Глава 6. КЛАССИФИКАЦИЯ БЕЗ ОБУЧЕНИЯ (ПАРАМЕТРИЧЕСКИЙ СЛУЧАЙ). РАСЩЕПЛЕНИЕ СМЕСЕЙ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ

6.1. Понятие смеси вероятностных распределений

6.1.1. Примеры. Начнем пояснение понятия смеси распределений с рассмотрения ряда конкретных примеров.

Пример 6.1. Контроль (по количественному признаку) изделий (полуфабрикатов) совокупности, составленной из продукции двух разных станков. В отдел технического контроля (ОТК) поступают партии изделий, составленные с помощью случайного извлечения из объединенной продукции двух станков (станка A и станка B). Изделия контролируются по некоторому количественному параметру (линейному размеру) ξ мм, так что результатом контроля i-го изделия партии является число x_i мм (изделия на станках не маркируются, так что в ОТК не известно, на каком именно станке произведено каждое из них). Производительность станка A в 1,5 раза выше производительности станка B. Задано номинальное значение контролируемого параметра $a = 65$ мм и известно, что точность работы станков характеризуется одинаковой величиной среднеквадратических отклонений $\sigma_A = \sqrt{\mathbf{D}\xi_A}$ и $\sigma_B = \sqrt{\mathbf{D}\xi_B}$, равной 1,0 мм[1]. Позже выяснилось, что станок A был настроен правильно (производил изделие со средним значением $\mathbf{E}\xi_A = 65$ мм, равным номиналу), в то время как настройка станка B была сбита в направлении завышения номинала (а именно $\mathbf{E}\xi_B = 67$ мм).

Известно также, что распределение размеров изделий, произведенных на каком-то определенном станке, описывается *нормальным законом* с параметрами $a_\gamma = \mathbf{E}\xi_\gamma$ и $\sigma_\gamma^2 = \mathbf{D}\xi_\gamma$ ($\gamma = A$ или $\gamma = B$).

Очевидно, анализируемая в ОТК по наблюдениям $x_1, x_2, \ldots, x_n \ldots$ генеральная совокупность будет состоять из *смеси двух нормальных генеральных совокупностей*, одна из

[1] Случайная величина ξ будет снабжаться нижним индексом (A или B) в тех случаях, когда речь идет о продукции какого-то определенного станка (соответственно станка A или станка B).

которых представляет продукцию станка A и описывается в соответствии с вышесказанным плотностью

$$f_A(x) = \varphi(x;\ a_A,\ \sigma_A^2) = \frac{1}{\sqrt{2\pi\sigma_A^2}}\, e^{\frac{-(x-a_A)^2}{2\sigma_A^2}},$$

а другая — продукцию станка B и описывается плотностью

$$f_B(x) = \varphi(x;\ \ a_B,\ \sigma_B^2) = \frac{1}{\sqrt{2\pi\sigma_B^2}}\, e^{\frac{-(x-a_B)^2}{2\sigma_B^2}}.$$

Обозначая $\theta_\gamma = (a_\gamma,\ \sigma_\gamma^2)$, а удельный вес изделий станка γ через P_γ ($\gamma = A,\ B$), можем записать уравнение функции плотности $f(x)$, описывающей закон распределения анализируемого признака ξ *во всей* (*объединенной*) генеральной совокупности, в виде:

$$f(x) = p_A\, \varphi(x;\ \theta_A) + p_B\, \varphi(x;\ \theta_B). \qquad (6.1)$$

Учитывая, что в объединенной генеральной совокупности продукции станка A в 1,5 раза больше, чем продукции станка B (поскольку производительность станка A в 1,5 раза выше), а также то, что $a_A = 65$ мм, $a_B = 67$ мм, $\sigma_A^2 = \sigma_B^2 = 1$ мм^2, имеем:

$$f(x) = 0{,}6\, \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(x-65)^2}{2}} + 0{,}4\, \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(x-67)^2}{2}}. \qquad (6.1')$$

Правыми частями уравнений (6.1) и (6.1′) и представлен частный случай того, что принято называть *смесью вероятностных распределений* [1].

На рис. 6.1 представлены графики фуикций плотности $f_A(x)$, $f_B(x)$ и $f(x)$.

В соотношениях (6.1) и (6.1′) величины $p_A = 0{,}6$ и $p_B = 0{,}4$ представляют удельные веса соответствующих компонентов смеси (их еще называют *априорными вероятностями* появления наблюдений именно из данного компонента смеси), а $\theta_A = (a_A,\ \sigma_A^2)$ и $\theta_B = (a_B,\ \sigma_B^2)$ — векторные параметры, от значений которых зависят законы распределения компонентов смеси.

---

[1] Речь идет о *частном случае*, поскольку в общей модели смеси распределений, во-первых, могут участвовать более чем два (и даже континуум) составляющих смесь распределения, а во-вторых, анализируемые распределения могут быть многомерными и не обязаны быть однотипными (в данном примере оба компонента — нормальные).

Если сотрудники ОТК или потребители изделий-полуфабрикатов захотят по наблюдениям $x_1, x_2, \ldots$ определить, *на каком именно станке произведено каждое из них*, то как раз и возникает одна из типичных задач классификации наблюдений в условиях отсутствия обучающих выборок (конечно, в данном примере можно представить себе специально организованное производство этих изделий, в результате

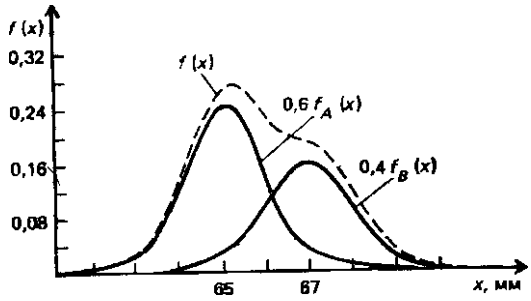


Рис. 6.1. Графики функции плотности отдельных компонентов и самой смеси из примера 6.1
— — — — для $f(x)$

которого можно получнть отдельно изделия от станка A и отдельно — от станка B и использовать их в дальнейшем в качестве обучающих выборок).

Пример 6.2. Выявление и анализ типов потребительского поведения по данным обследований семейных бюджетов [154, с. 47—48, 68—69]. Здесь представлен один из фрагментов исследования, проведенного с целью изучения (на базе семейных бюджетов) дифференциации потребностей, выявления основных типов потребительского поведения и определения главных типообразующих признаков (социально-демографической, региональной, экономической природы). Исследуемым многомерным признаком является вектор Y показателей $y^{(1)}$, $y^{(2)}$, ..., $y^{(p)}$ потребительского поведения семьи, т. е. каждой (i-й) обследованной семье ставится в соответствие многомерное наблюдение

$$Y_i = \begin{pmatrix} y_i^{(1)} \\ \vdots \\ y_i^{(p)} \end{pmatrix}, \quad i = 1, 2, \ldots, n,$$

где $y_{(i)}^{(m)}$ — удельное (т. е. рассчитанное в среднем на одного члена семьи) количество m-го вида благ (товаров или услуг,

включая сбережения), потребляемое i-й обследованной семьей в базовый период (за год) и выраженное в натуральных или денежных единицах.

В соответствии с одним из принятых в исследовании базовых исходных допущений постулируется существование в анализируемом пространстве $\mathbf{\Pi}^p(Y)$ $(Y \in \mathbf{\Pi}^p(Y))$ сравнительно небольшого (и неизвестного) числа k типов потребительского поведения, таких, что различия в структуре потребления Y семей одного типа носят *случайный характер* (т. е. обусловлены влиянием множества случайных, не поддающихся управлению и учету факторов) *и незначительны* по сравнению с различиями в потребительском поведении семей, представляющих разные типы. При этом предполагается, что случайный разброс структур потребительских поведений $Y(j)$ *внутри* любого (j-го) типа описывается многомерным (в нашем случае p-мерным) *нормальным* законом распределения с некоторым вектором средних (и в то же время — наиболее характерных, наиболее часто наблюдаемых) значений

$$\mathbf{a}(j) = \begin{pmatrix} a_{(j)}^{(1)} \\ a_{(j)}^{(2)} \\ \vdots \\ a_{(j)}^{(p)} \end{pmatrix}$$

и с ковариационной матрицей

$$\mathbf{\Sigma}(j) = \begin{pmatrix} \sigma_{11}(j) & \sigma_{12}(j) & \dots & \sigma_{1p}(j) \\ \sigma_{21}(j) & \sigma_{22}(j) & \dots & \sigma_{2p}(j) \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1}(j) & \sigma_{p2}(j) & \dots & \sigma_{pp}(j) \end{pmatrix}.$$

(см. сведения о многомерном нормальном законе в [11, п. 6.1.5]).

Однако в начале исследования нет сведений об упомянутых гипотетических типах потребительского поведения: неизвестно ни их число k, ни значения определяющих эти типы многомерных параметров $\theta_j = (\mathbf{a}(j), \mathbf{\Sigma}(j))$. Поэтому вынуждены рассматривать имеющиеся в нашем распоряжении результаты бюджетных обследований семей

$$Y_1, Y_2, \dots, Y_n \qquad (6.2)$$

как выборку из генеральной совокупности, являющейся смесью многомерных нормальных законов распределения. Другими словами, функция плотности $f(Y)$, описывающая распреде-

ление вектора Y в этой объединенной генеральной совокупности, имеет вид

$$f(Y) = \sum_{j=1}^{k} p_j \varphi(Y;\ \mathbf{a}(j);\ \Sigma(j)), \tag{6.3}$$

где p_j $(j = 1, 2, \ldots, k)$ — не известный нам удельный вес (априорная вероятность) семей j-го типа потребительского поведения в общей совокупности семей;

$$\varphi(Y;\ \mathbf{a}(j);\ \Sigma(j)) =$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma(j)|^{\frac{1}{2}}} e^{-\frac{1}{2}(Y-\mathbf{a}(j))' \Sigma_{(j)}^{-1} (Y-\mathbf{a}(j))} —$$

многомерная нормальная плотность, описывающая закон распределения исследуемого признака $Y(j)$ внутри совокупности семей j-го типа потребительского поведения $(j = 1, 2, \ldots k)$.

Далее необходимо по выборке (6.2) оценить неизвестные значения параметров k, p_1, p_2, ..., p_{k-1}, $\mathbf{a}(j)$ и $\Sigma(j)$ $(j = 1, \ldots, k)$ модели (6.3), чтобы в конечном счете суметь расклассифицировать (в определенном смысле наилучшим образом) семьи (6.2) по искомым типам потребительского поведения. Общая схема действий, увязывающая задачу статистического оценивания параметров смеси типа (6.3) с задачей автоматической классификации, изложена в п.6.2.

6.1.2. Общая математическая модель смеси распределений. Рассмотренные в примерах смеси (6.1) и (6.3) представляют собой частные случаи общей модели смеси, определение которой дадим здесь. Обобщение рассмотренных в примерах смесей может быть произведено в направлении: 1) отказа от конечности и даже дискретности компонентов, составляющих смесь, распространения понятия смеси на непрерывную смешивающую функцию; 2) отказа от однотипности участвующих в смеси компонентов (под однотипностью компонентов-распределений понимается их принадлежность к общему параметрическому семейству распределений, например к нормальному).

Итак, пусть имеется двухпараметрическое семейство p-мерных плотностей (полигонов вероятностей) распределения

$$\mathbf{F} = \{f_\omega(X;\ \theta(\omega))\}, \tag{6.4}$$

где одномерный (целочисленный или непрерывный) параметр ω в качестве нижнего индекса функции f определяет

специфику общего вида каждого компонента — распределения смеси, а в качестве аргумента при многомерном, вообще говоря, параметре θ определяет зависимость значений хотя бы части компонентов этого параметра от того, в каком именно составляющем распределении f_ω он присутствует. И пусть

$$\mathbf{P} = \{P(\omega)\} \tag{6.5}$$

— семейство *смешивающих функций распределения.*

Функция плотности (полигон вероятностей) распределения

$$f(X) = \int f_\omega(X;\ \theta(\omega))\, dP(\omega)) \tag{6.6}$$

называется P-смесью (или просто смесью) распределений семейства $\mathbf{F}$ (интеграл в (6.6) понимается в смысле Лебега—Стильтьеса; см., например, [86]).

Нас интересует использование моделей смесей в теории и практике автоматической классификации, поэтому сузим данное выше определение смеси и будем рассматривать в дальнейшем лишь *случай конечного числа k* возможных значений параметра ω, что соответствует конечному числу скачков смешивающих функций $P(\omega)$. Величины этих скачков как раз и будут играть роль удельных весов (априорных вероятностей) p_j компонентов смеси ($j = 1, 2, \ldots, k$), так что (6.6) в этом случае может быть записано в виде

$$f(X) = \sum_{j=1}^{k} p_j f_j(X;\ \theta(j)). \tag{6.6'}$$

Если же дополнительно постулировать *однотипность* компонентов-распределений $f_j(X;\ \theta(j))$, т. е. принадлежность всех $f_j(X;\ \theta(j))$ к одному общему семейству $\{\tilde{f}(X;\ \theta)\}$, то модель смеси может быть представлена в виде

$$f(X) = \sum_{j=1}^{k} p_j \tilde{f}(X;\ \theta(j)). \tag{6.6''}$$

Интерпретация в задачах автоматической классификации j-го компонента смеси (j-й генеральной совокупности) в качестве j-го искомого класса (сгустка, скопления) обусловливает естественность дополнительного ограничения условия, накладываемого на плотности (полигоны вероятностей) $f_j(X;\ \theta(j))$ и заключающегося в *их одномодальности.*

6.1.3. Задача расщепления смеси распределений. Решить

эту задачу в выборочном варианте — значит по выборке классифицируемых наблюдений

$$X_1, X_2, \ldots, X_n, \tag{6.7}$$

извлеченной из генеральной совокупности, являющейся смесью (6.6) генеральных совокупностей типа (6.4) (при заданном общем виде составляющих смесь функций $f_\omega(X, \theta(\omega))$, построить статистические оценки для числа компонентов смеси k, их удельных весов (априорных вероятностей) p_1, p_2, ..., p_k и, главное, для каждого из компонентов $f_\omega(X; \theta(\omega))$ анализируемой смеси (6.6). В некоторых частных случаях имеющиеся априорные сведения дают исследователю точное знание числа компонентов смеси k, а иногда и априорных вероятностей p_1, p_2, ..., p_k. Тогда задача расщепления смеси сводится лишь к оцениванию функций $f_\omega(X; \theta(\omega))$.

Однако не следует ставить знак тождества между задачей расщепления смеси и задачей статистического оценивания параметров в модели (6.6) по выборке (6.7), поскольку задача расщепления сохраняет смысл и *применительно к генеральным совокупностям*, т. е. в теоретическом варианте. В этом случае она заключается в восстановлении компонентов $f_\omega(X; \theta(\omega))$ и смешивающей функции $P(\omega)$ по заданной левой части $f(X)$ соотношения (6.6) и называется задачей идентификации компонентов смеси. В п. 6.3 показано, что эта задача не всегда имеет единственное решение.

6.2. Общая схема решения задачи автоматической классификации в рамках модели смеси распределений (сведение к схеме дискриминантного анализа)

Базовая идея, лежащая в основе принятия решения, к какой из k анализируемых генеральных совокупностей отнести данное классифицируемое наблюдение X_i, состоит в том, что *наблюдение следует отнести к той генеральной совокупности, в рамках которой оно выглядит наиболее правдоподобным*. Другими словами, если дано точное описание (например, в виде функций $f_1(X)$, ..., $f_k(X)$ плотности в непрерывном случае или полигонов вероятностей в дискретном) конкурирующих генеральных совокупностей, то следует поочередно вычислить значения функций правдоподобия для данного наблюдения X_i в рамках каждой из рассматриваемых генеральных совокупностей (т. е. вычислить значения $f_1(X_i)$, $f_2(X_i)$, ..., $f_k(X_i)$) и отнести X_i к тому классу,

функция правдоподобия которого максимальна[1]. Если же известен лишь общий вид функций $f_1(X; \theta_1)$, $f_2(X; \theta_2)$, ..., $f_k(X; \theta_k)$, описывающих анализируемые классы, но не известны значения, вообще говоря, многомерных параметров $\theta_1, \theta_2, \ldots, \theta_k$, и если при этом располагают так называемыми обучающими выборками, то данный случай лежит в рамках параметрической схемы дискриминантного анализа и порядок действий будет следующим (см. гл. 2 и 3): сначала по j-й обучающей выборке оцениваем параметр θ_j ($j = 1, 2, \ldots, k$), а затем производим классификацию наблюдений, руководствуясь тем же самым принципом максимального правдоподобия, что и в случае полностью известных функций $f_j(X)$.

В схеме автоматической классификации, опирающейся на модель смеси распределений, как и в схеме параметрического ДА, задающие искомые классы функции $f_1(X; \theta_1)$, $f_2(X; \theta_2)$,... также известны лишь с точностью до значений параметров. Но в схеме автоматической классификации неизвестные значения параметров $\theta_1, \theta_2, \ldots$, так же, впрочем, как и параметров $k, p_1, p_2, \ldots, p_k$, *оцениваются не по обучающим выборкам* (*их нет в распоряжении исследователя*), *а по классифицируемым наблюдениям* $X_1, X_2, \ldots, X_n$ с помощью одного из известных методов статистического оценивания параметров (метода максимального правдоподобия, метода моментов или какого-либо другого; см. о процедурах статистического оценивания параметров смеси в § 6.4). Начиная с момента, когда по выборке (6.7) сумели получить оценки $\widehat{k}$, $\widehat{p_1}, \widehat{p_2}, \ldots, \widehat{p_k}, \widehat{\theta_1}, \ldots, \widehat{\theta_k}$ неизвестных параметров $k, p_1, p_2, \ldots, p_k, \theta_1, \theta_2, \ldots, \theta_k$ модели (6.6′) или (6.6″), снова имеем схему дискриминантного анализа и собственно процесс классификации наблюдений (6.7) производим точно так же, как и в схеме параметрического ДА (т. е. относим наблюдение X_i к классу с номером j_0, если $p_{j_0} f_{j_0}(X_i; \widehat{\theta}_{j_0}) = \max\limits_{1 \leqslant j \leqslant k} \{p_j \times \times f_j(X_i; \widehat{\theta}_j)\}$).

Итак, *главное отличие схемы параметрического ДА от схемы автоматической классификации, производимой в рамках модели смеси распределений,—в способе оценивания неизвестных параметров, от которых зависят функции, описывающие классы.* Но оценивание параметров в модели смеси—процесс неизмеримо более сложный, чем оценивание параметров по обучающим выборкам.

---

[1] Для большей ясности здесь подразумевается простой случай равных априорных вероятностей и равных потерь от неправильного отнесения наблюдения X_i к любому из классов. Более общая схема и более подробно представлена в гл. 1.

6.3. Идентифицируемость (различимость) смесей распределений

Семейство смесей (6.6) ($P(\omega) \in \mathbf{P}$, см. (6.5)) называется *идентифицируемым* (различимым), если из равенства

$$\int f_\omega(X;\ \theta(\omega))\,\mathrm{d}P(\omega) = \int f_\omega(X;\ \theta(\omega))\,\mathrm{d}P^*(\omega)$$

следует, что $P(\omega) \equiv P^*(\omega)$ для всех $P(\omega) \in \mathbf{P}$.

Поскольку нас интересуют в первую очередь *конечные* смеси типа (6.6′), переформулируем понятие идентифицируемости (различимости) смесей специально применительно к ним.

Конечная смесь (6.6′) называется *идентифицируемой* (различимой), если из равенства

$$\sum_{j=1}^{k} p_j f_j(X;\ \theta(j)) = \sum_{l=1}^{k^*} p_l^* f_l(X;\ \theta^*(l))$$

следует: $k = k^*$ и для любого j $(1 \leqslant j \leqslant k)$ найдется такое l $(1 \leqslant l \leqslant k^*)$, что $p_j = p_l^*$ и $f_j(X;\ \theta(j)) \equiv f_l(X;\ \theta^*(l))$.

В работах [320, 321, 327] сформулированы необходимые и достаточные условия различимости для непрерывных и конечных смесей. Из них, в частности, следует, что различными являются конечные смеси из распределений: 1) нормальных (в том числе многомерных); 2) экспоненциальных; 3) пуассоновских; 4) Коши. Описание и свойства перечисленных распределений см., например, в [11, гл. 6]. В то же время конечные смеси биномиальных, равномерных распределений в общем случае не являются идентифицируемыми. При определенном классе смешивающих распределений не являются идентифицируемыми и непрерывные смеси нормальных распределений. Поясним это на примерах.

Пример 6.3. Пусть семейство компонентов смеси состоит из *равномерных* распределений с неизвестными параметрами, т. е. $\theta = (a, \sigma)$ и плотность

$$f(X;\ \theta) = f(X;\ a,\ \sigma) = \begin{cases} 0 \text{ при } X > a+\sigma, \\ \dfrac{1}{2\sigma} \text{ при } a-\sigma \leqslant X \leqslant a+\sigma, \\ 0 \text{ при } X < a-\sigma. \end{cases}$$

Рассмотрим класс конечных смесей, когда функция $P(\omega) = P(a)$ имеет лишь два скачка, что соответствует

смешиванию двух различных однородных классов. Легко проверить, что для любого λ $(0 \leqslant \lambda \leqslant 1)$

$$f(X;\ a,\ \sigma) = \lambda f(X;\ a - \sigma(1-\lambda),\ \sigma\lambda) + (1-\lambda) f(X;\ a + \sigma\lambda,\ \sigma(1-\lambda)).$$

Это означает, что смешивающая функция $P_\lambda(a)$ делает два скачка величины λ и $(1-\lambda)$ и если $\lambda_1 \neq \lambda_2$, то $P_{\lambda_1}(a) \neq P_{\lambda_2}(a)$. Аналогично можно произвести разбиение для любого числа классов. На рис. 6.2 представлен частный слу-

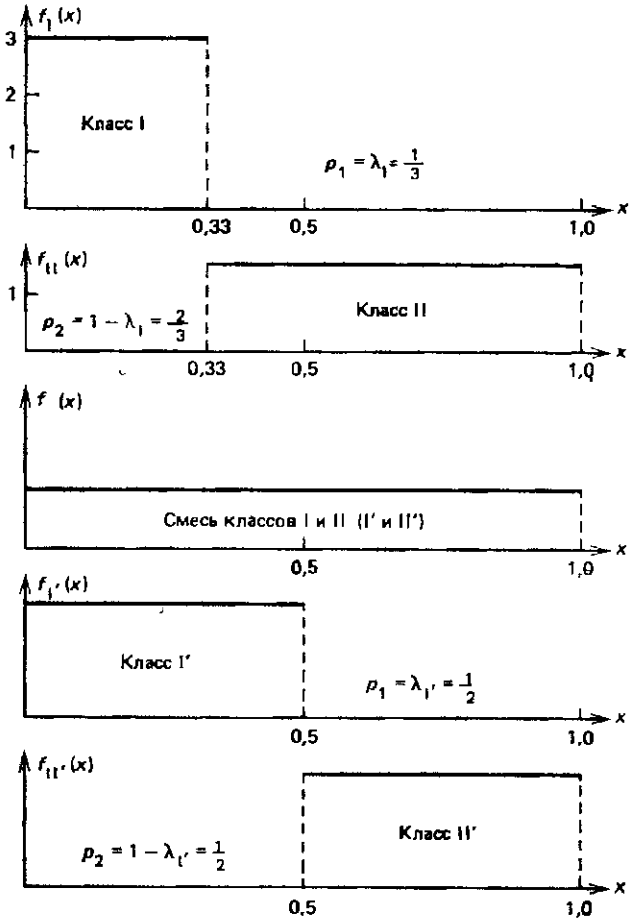


Рис. 6.2. Пример неразличимых смесей: произвольное разбиение точек, равномерно распределенных на отрезке прямой, на два класса

чай разбиения на два класса, когда два разных варианта смешивающих функций (1-й вариант: $\lambda_I = \frac{1}{3}$, 2-й вариант: $\lambda_{I'} = \frac{1}{2}$) приводят к одному и тому же выражению для плотности смеси распределенний. Другими словами, однородная группа представителей, которые могут появиться равновероятно в любой точке неопределенной области, может трактоваться как смесь (даже конечная) групп представителей, однородных в том же смысле. Но если об области, где могут появляться представители, кое-что известно, например в данном случае $\sigma = \frac{1}{2}$, то равномерное распределение уже нельзя разбить на смесь двух равномерных распределений с $\sigma = \frac{1}{2}$.

Пример 6.4. Рассмотрим семейство двумерных равномерных распределений на секторах круга единичного радиуса с центром в точке (0,0). Сектор задается начальным направлением φ и углом при вершине $\beta > 0$, т. е. $\mathbf{F} = \{f(X;\ \theta)\}$, где $X = (x^{(1)},\ x^{(2)})$, $a\ \theta = (\varphi,\ \beta)$. Таким образом, для любых φ, β_1, β_2 $(\beta_1 + \beta_2 \leqslant 2\pi)$ выполняется равенство

$$\frac{\beta_1}{\beta_1 + \beta_2} f(X;\ \varphi,\ \beta_1) + \frac{\beta_2}{\beta_1 + \beta_2} f(X;\ \varphi + \beta_1, \beta_2) =$$
$$= f(X;\ \varphi,\ \beta_1 + \beta_2),$$

что означает, что семейство смесей $\mathbf{F}$ неразличимо. Следовательно, равномерное распределение на круге с плотностью $f(X;\ 0,\ 2\pi) \in \mathbf{F}$ можно представить в виде

$$\frac{1}{2} f(X;\ \varphi,\ \pi) + \frac{1}{2} f(X;\ \varphi + \pi,\ \pi).$$

Это означает, что возможно любое разделение точек на два класса прямой, проходящей через центр.

6.4. Процедуры оценивания параметров модели смеси распределений

Итак, из § 6.2 известно, что задача автоматической классификации многомерных наблюдений (6.7), решаемая в рамках модели смеси распределений вида (6.6′), может быть сведена к обычной схеме дискриминантного анализа: необходимым предварительным этапом этой редукции является процесс статистического оценивания по выборке (6.7) (которую будем полагать в дальнейшем *случайной*, состоя-

щей из n *независимых* наблюдений многомерного признака X с законом распределения (6.6′)) неизвестных параметров $k, p_1, p_2, \ldots, p_{k-1}, \theta_1, \theta_2, \ldots, \theta_k$. Во всем дальнейшем изложении материала данной главы предполагается, что анализируемая смесь идентифицируема (различима).

И в теоретико-методическом, и в вычислительном плане проблема построения и анализа свойств процедур оценивания параметров смесей вида (6.6′) по выборке (6.7) является весьма сложной. Одна из главных трудностей связана с оцениванием целочисленного параметра k — числа компонентов (или числа классов) анализируемой смеси. Во всех описываемых ниже процедурах (кроме процедуры SEM) схема оценивания строится таким образом, что вначале заготавливаются оценки параметров p_j и θ_j ($j = 1, 2, \ldots, k$) для последовательности *фиксированных* значений k ($k = 1, 2, \ldots, K$, где K — некоторая гарантированная мажоранта для возможного числа классов), а затем с помощью того или иного приема подбирается «наилучшее» значение k в качестве оценки для не известного нам истинного числа классов k_0.

6.4.1. Процедуры, базирующиеся на методе максимального правдоподобия. В данном пункте речь идет о процедурах, позволяющих находить максимум (по параметрам $p_1, p_2, \ldots, p_{k-1}, \theta_1, \theta_2, \ldots, \theta_k$ при фиксированном k) определяемой с помощью соотношения (6.6) логарифмической функции правдоподобия (о функции правдоподобия см. [11, § 8.2]), т. е. о решении оптимизационной задачи вида

$$\sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} p_j f(X_i;\ \theta_j) \right) \to \max_{p_j,\ \theta_j}. \tag{6.8}$$

Наиболее работоспособная общая схема построения процедур, позволяющих находить решения задачи (6.8), была впервые, по-видимому, предложена в работах [166, 209, 210], а затем развита в [333, 212, 254, 295] и др. Конкретные алгоритмы, построенные по этой схеме, часто называют *алгоритмами типа* EM, поскольку в каждом из них можно выделить два этапа, находящихся по отношению друг к другу в последовательности итерационного взаимодействия: оценивание (*Estimation*) и максимизация (*Maximisation*).

Общая схема построения процедур и их некоторые свойства. Введем в рассмотрение так называемые *апостериорные вероятности* g_{ij} принадлежности наблюдения X_i к j-му классу:

$$g_{ij} = \frac{p_j f(X_i;\ \theta_j)}{\sum_{j=1}^{k} p_j f(X_i;\ \theta_j)} \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, k). \tag{6.9}$$

Очевидно, $g_{ij} \geqslant 0$ и $\sum_{j=1}^{k} g_{ij} = 1$ для всех $i = 1, 2, \ldots, n$. Затем обозначим $\Theta = (p_1, p_2, \ldots, p_k; \theta_1, \theta_2, \ldots, \theta_k)$ и представим анализируемую логарифмическую функцию правдоподобия

$$\ln L(\Theta) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} p_j f(X_i; \theta_j) \right) \text{ в виде}$$

$$\ln L(\Theta) = \sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} \ln p_j + \sum_{j=1}^{k} \sum_{i=1}^{n} q_{ij} \ln f(X_i; \theta_j) -$$

$$- \sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} \ln g_{ij} \qquad (6.10)$$

(справедливость этого тождества легко проверяется с учетом (6.9) и того, что $\sum_{j=1}^{k} g_{ij} = 1$).

Далее идея построения итерационного алгоритма вычисления оценок $\widehat{\Theta} = (\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_k; \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k)$ для параметров $\Theta = (p_1, p_2, \ldots, p_k; \theta_1, \theta_2, \ldots, \theta_k)$ состоит в том, что, отправляясь от некоторого начального приближения $\widehat{\Theta}^0$, вычисляют (по формулам (6.9)) начальные приближения g_{ij}^0 для апостериорных вероятностей g_{ij} (*этап оценивания*), а затем, возвращаясь к (6.10), при вычисленных значениях g_{ij}^0 определяют значения $\widehat{\Theta}^1$ из условия максимизации *отдельно* каждого из первых двух слагаемых правой части (6.10) (*этап максимизации*), поскольку первое слагаемое $(\sum_{j=1}^{k} \sum_{i=1} g_{ij} \ln p_j)$ зависит только от параметров p_j $(j = 1, 2, \ldots, k)$, а второе слагаемое $(\sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} \ln f(X_i; \theta_j))$ зависит только от параметров θ_j $(j = 1, 2, \ldots, k)$.

Очевидно, решение оптимизационной задачи

$$\sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij}^{(t)} \ln p_j \to \max_{p_1, p_2, \ldots, p_k} \qquad (6.11)$$

дается выражением (с учетом $\sum_{j=1}^{k} p_j = 1$)

$$p_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} g_{ij}^{(t)}, \qquad (6.12)$$

здесь t — номер итерации, $t = 0, 1, 2, \ldots$

Решение оптимизационной задачи

$$\sum_{j=1}^{k}\sum_{i=1}^{n} g_{ij}^{(t)} \ln f(X_i;\ \theta_j) \to \max_{\theta_1,\ \theta_2,\ \ldots,\ \theta_k} \tag{6.13}$$

получить намного проще решения задачи (6.8): выражение для $\theta_j^{(t+1)}$ записывается с учетом знания конкретного вида функций $f(X;\ \theta)$. Ниже приведены выражения для $\theta_j^{(t+1)}$ (при заданных $g_{ij}^{(t)}$) для случая *нормальных* плотностей $f(X;\ \theta)$.

В той же работе М. И. Шлезингера, где эта схема (позднее названная EM-схемой) впервые предложена [166], установлены и основные свойства реализующих ее алгоритмов (позднее в работах [334, 197, 295, 222] эти свойства были передоказаны и частично развиты). В частности, было доказано, что при достаточно широких предположениях (наиболее неприятным, жестким из них является требование ограниченности логарифмической функции правдоподобия, которое, правда, было неправомерно опущено в формулировках [166]) предельные точки всякой последовательности, порожденной итерациями EM-алгоритма, являются стационарными точками оптимизируемой логарифмической функции правдоподобия $\ln L(\Theta)$ и что найдется неподвижная точка алгоритма, к которой будет сходиться каждая из таких последовательностей. Если дополнительно потребовать положительной определенности информационной матрицы Фишера для $\ln L(\Theta)$ при истинных значениях параметра Θ [11, § 8.2], то можно показать [295], что асимптотически по $n \to \infty$ (т. е. при больших выборках (6.7)) существует единственное сходящееся (по вероятности) решение $\widehat{\Theta}(n)$ уравнений метода максимального правдоподобия и, кроме того, существует в пространстве параметров Θ норма, в которой последовательность $\Theta^{(t)}(n)$, порожденная EM-алгоритмом, сходится к $\widehat{\Theta}(n)$, если только начальная аппроксимация Θ^0 не была слишком далека от $\widehat{\Theta}(n)$.

Таким образом, результаты исследования свойств EM-алгоритмов метода максимального правдоподобия расщепления смеси и их практическое использование показали, что они являются достаточно работоспособными (при известном числе компонентов смеси) даже при большом числе k компонентов и при высоких размерностях p анализируемого признака X.

Основными «узкими местами» этого подхода являются: необходимость предъявления *требования ограниченности*

к анализируемой функции правдоподобия $L(\Theta)$, высокая сложность и трудоемкость процесса вычислительной реализации соответствующих процедур и медленная сходимость порождаемых ими итерационных процессов.

Смеси нормальных классов. Продолжим исследование задачи статистического оценивания параметров Θ смеси (6.6″), состоящей из известного числа k классов. Дополнительно постулируем при этом, что каждый объект X класса j представляет собой элемент *нормальной* генеральной совокупности $N(\mathbf{a}_j, \Sigma)$, где векторы средних $\mathbf{a}_j$ различны для разных классов, а ковариационные параметры Σ совпадают, но неизвестны компоненты ни $\mathbf{a}_j$ $(j = 1, 2, \ldots, k)$, ни Σ. Кроме того, неизвестны априорные вероятности классов p_j $(j = 1, 2, \ldots, k)$.

Легко проверить [210], что в этом случае

$$g_{ij} = \frac{\exp[\alpha_j' X_i + \beta_j]}{\sum_{j=1}^{k} \exp[\alpha_j' X_i + \beta_j]},$$

где

$$\alpha_j = \Sigma^{-1}\mathbf{a}_j \text{ и } \beta_j = \frac{1}{2}\mathbf{a}_j' \Sigma^{-1}\mathbf{a}_j + \ln p_j.$$

Учитывая описанную выше схему EM-алгоритма, следует определить процедуру, которая максимизировала бы

$$\ln L(X_1, \ldots, X_n; \Theta_j) = \ln L_j = \sum_{i=1}^{n} g_{ij}^{(t)} \ln\left[\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \times \right.$$

$$\left. \times e^{-\frac{1}{2}(X_i - \mathbf{a}_j)' \Sigma^{-1} (X_i - \mathbf{a}_j)}\right]$$

по $\mathbf{a}_j$ и Σ, или, учитывая, что в данном случае $\Theta_j = (\mathbf{a}_j, \Sigma)$, определить процедуру, которая максимизировала бы

$$\sum_{j=1}^{k} \ln L_j = \sum_{j=1}^{k}\sum_{i=1}^{n} g_{ij}^{(t)} \ln\left[\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \times \right.$$

$$\left. \times e^{-\frac{1}{2}(X_i - \mathbf{a}_j)' \Sigma^{-1} (X_i - \mathbf{a}_j)}\right]$$

при условии, что $g_{ij}^{(t)}$ каким-либо способом уже получены. Эта процедура даст величины $\theta_1^{(t+1)} = \mathbf{a}_j^{(t+1)}$ для $(t+1)$-го шага и $\theta_2^{(t+1)} = \Sigma^{(t+1)}$ по данным $\theta_1^{(t)}$ и $\theta_2^{(t)}$. Два последую-

щих утверждения определяют точку максимума для $\ln L_j$ и $\sum_{j=1}^{k} \ln L_j$ в итерационной процедуре, построенной по схеме ЕМ-алгоритма.

Для простоты их формулировки будем опускать индекс t, подчеркивающий связь с шагом процедуры. Напомним, что последовательность g_{ij} $(j=1, 2, \ldots, k;\ i=1, 2, \ldots, n)$ такова, что

$$g_{ij} \geqslant 0, \quad \sum_{i=1}^{n} g_{ij} = g_{\cdot j} > 0, \quad \sum_{i=1}^{n} \sum_{j=1}^{k} g_{ij} = n.$$

Утверждение 1. Пусть g_{ij} — определенная выше последовательность и $f(X \mid \theta_j)$ — p-мерные нормальные плотности, такие, что $\theta_j = (a_j, \Sigma_j)$. Тогда для любых вектор-столбцов $X_1, X_2, \ldots, X_n$ величины $\ln L_j$ $(j = 1,2, \ldots, k)$ достигают максимума при

$$\widehat{a}_j = \frac{1}{g_{\cdot j}} \sum_{i=1}^{n} g_{ij} X_i,$$

$$\widehat{\Sigma}_j = \frac{1}{g_{\cdot j}} \sum_{i=1}^{n} g_{ij} (X_i - \widehat{a}_j)(X_i - a_j)'.$$

Утверждение 2. Пусть g_{ij} — определенная выше последовательность и $f(X \mid \theta_j)$ — p-мерные нормальные плотности, такие, что $\theta_j = (a_j, \Sigma)$. Тогда для любых вектор-столбцов $X_1, X_2, \ldots, X_n$ величина $\sum_{j=1}^{k} \ln L_j$ достигает максимума при

$$\widehat{a}_j = \frac{1}{g_{\cdot j}} \sum_{i=1}^{n} g_{ij} X_i, \quad \widehat{\Sigma} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} (X_i - \widehat{a}_j)(X_i - \widehat{a}_j)'$$

и

$$\max_{a_j, \Sigma} \sum_{j=1}^{k} \ln L_j = -\frac{np}{2} |\ln (2\pi)| - \frac{n}{2} \ln |\widehat{\Sigma}|.$$

Доказательство этих утверждений опирается на леммы 3.2.1 и 3.2.2 из [16].

Таким образом, при заданных

$$g_{ij}^{(t)} = \frac{\exp [\alpha_j'(t) X_i + \beta_j(t)]}{\sum_{j=1}^{k} \exp [\alpha_j'(t) X_i + \beta_j(t)]},$$

где

$$\alpha_j(t) = (\Sigma^{(t)})^{-1}\,\widehat{\mathbf{a}}_j^{(t)} \quad \text{и} \quad \beta_j(t) = \frac{1}{2}\,\widehat{\mathbf{a}}_j^{(t)\prime}\,(\Sigma^{(t)})^{-1}\,\widehat{\mathbf{a}}_j^{(t)} + \ln p_j^{(t)},$$

величины

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} g_{ij}^{(t)}\left(X_i - \widehat{\mathbf{a}}_j^{(t)}\right)\left(X_i - \widehat{\mathbf{a}}_j^{(t)}\right)'$$

и

$$\widehat{\mathbf{a}}_j^{(t)} = \frac{\sum_{i=1}^{n} g_{ij}^{(t)} X_i}{\sum_{j=1}^{n} g_{ij}^{(t)}}$$

максимизируют $\sum_{j=1}^{k} \ln L_j$.

Далее легко получить, что

$$\widehat{p}_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} g_{ij}^{(t)}$$

и

$$\Theta^{(t+1)} = \left(p_j^{(t+1)},\ \mathbf{a}_j^{(t+1)},\ \Sigma^{(t+1)};\ j = 1,\ 2,\ \ldots,\ k\right).$$

Если существуют пределы

$$\lim_{t\to\infty} p_j^{(t)} = \widehat{p}_j,\ \lim_{t\to\infty} \widehat{\mathbf{a}}_j^{(t)} = \widehat{\mathbf{a}}_j,\ j = 1,\ 2,\ \ldots,\ k,\ \lim_{t\to\infty} \Sigma^{(t)} = \widehat{\Sigma},$$

то точка $\widehat{\Theta} = (\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_k, \widehat{\mathbf{a}}_1, \widehat{\mathbf{a}}_2, \ldots, \widehat{\mathbf{a}}_k, \widehat{\Sigma})$ является точкой максимума функции правдоподобия (возможно, правда, что этот максимум является *локальным*).

Легко видеть, что в качестве начальных данных можно задать не точку $\Theta^{(0)} = (p_1^{(0)}, \ldots, p_k^{(0)}, \mathbf{a}_1^{(0)}, \ldots, \mathbf{a}_k^{(0)}, \Sigma^{(0)})$, а набор величин $\alpha_j(0)$, $\beta_j(0)$, с помощью которых можно получить $g_{ij}^{(0)}$ и т. д. Именно такая итерационная процедура предлагается в работе [210].

Замечание. Точки, для которых $g_{ij} = 1/k$, являются *неподвижными* точками итерационной процедуры, но представляют собой посторонние точки, так как в этом случае $\mathbf{a}_j = \mathbf{a}$ $(j = 1,2, \ldots, k)$.

В случае двух классов $(k = 2)$, как показано в [203], процедура сильно упрощается. Для произвольных $\alpha'(0) = (\alpha_1(0), \ldots, \alpha_k(0))$ и $\beta(0)$, имеем

$$g_{i1}^{(0)} = \frac{1}{1 + \exp\left[\alpha'(0)\,X_i + \beta(0)\right]};\quad g_{i2}^{(0)} = 1 - g_{i1}^{(0)}.$$

$$\mathbf{a}_j^{(0)} = \frac{\sum_{i=1}^{n} g_{ij}^{(0)} X_i}{\sum_{i=1}^{n} g_{ij}^{(0)}}; \quad p^{(0)} = \frac{1}{n} \sum_{i=1}^{n} g_{ij}^{(0)}.$$

Далее определяются уточнения α и β следующим образом:

$$\alpha(1) = \frac{\mathbf{V}^{-1}(\mathbf{a}_2^{(0)} - \mathbf{a}_1^{(0)})}{1 - \pi^{(0)}(1 - \pi^{(0)})(\mathbf{a}_1^{(0)} - \mathbf{a}_2^{(0)})' \mathbf{V}^{-1}(\mathbf{a}_1^{(0)} - \mathbf{a}_2^{(0)})};$$

$$\beta(1) = -\frac{1}{2}\alpha'(0)(\mathbf{a}_1^{(0)} + \mathbf{a}_2^{(0)}) + \ln\frac{1-p^{(0)}}{p^{(0)}},$$

где

$$\mathbf{V} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})',$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Подставляя $\alpha(1)$ и $\beta(1)$ вместо $\alpha(0)$ и $\beta(0)$, можно итерационную процедуру продолжить до тех пор, пока значения α и β не перестанут изменяться. Далее, после того как значения $\widehat{\alpha}$ и $\widehat{\beta}$ установятся, можно определить оценку ковариационной матрицы

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}[(X_i - \widehat{\mathbf{a}_1})(X_i - \widehat{\mathbf{a}_1})' g_{i1} +$$

$$+ (X_i - \widehat{\mathbf{a}_2})(X_i - \widehat{\mathbf{a}_2})' g_{i2}].$$

Естественно точку X_i отнести к классу 1, если $g_{i1} > g_{i2}$, т. е. если $g_{i1} > 1/2$. Отсюда следует, что X_i будет отнесена к классу 1, если $\widehat{\alpha}' X_i + \widehat{\beta} < 0$, или к классу 2, если $\widehat{\alpha}' X_i + \widehat{\beta} > 0$. Следовательно, $\widehat{\alpha}' X + \widehat{\beta} = 0$ будет оценкой *разделяющей поверхности* классов 1 и 2, а $\widehat{\alpha}$ и $\widehat{\beta}$ — оценками параметров разделяющей поверхности (см. гл. 2,3).

Основные трудности этого метода классификации состоят в том, что скорость сходимости итерационного процесса зависит от расстояния Махаланобиса $\rho(\mathbf{a}_1, \mathbf{a}_2)$ между классами (см. гл. 1) и от начальных значений искомых параметров. Более того, может быть несколько локальных макси-

мумов и требуется, изменяя начальные данные, определить абсолютный максимум. Грубо говоря, итерационный процесс сходится к абсолютному максимуму $\widehat{\alpha}, \widehat{\beta}$ (при $k = 2$) из точек α (0), β (0), если угол между $\widehat{\alpha}$ и α (0) менее 45°. Это ясно показывает возрастание трудностей при росте размерности. Если точка α (0) выбрана случайно, то вероятность выполнения этого условия при $p = 5$ равна 0,076, при $p = 10 - 0{,}01$, при $p = 15 - 0{,}001$, при $p = 20 - 0{,}0002$. Поэтому при больших размерностях наблюдений ($p > 10$) желательно предварительно эту размерность снизить (например, методом главных компонент; см. раздел III).

Пример 6.5. Неограниченная функция правдоподобия. Рассмотрим простейший случай, когда число классов $k = 2$ и наблюдаемые величины X_i ($i = 1, 2, \ldots, n$) являются одиомерными ($p = 1$). Плотность распределения смеси

$$f(X) = p_1 \frac{1}{\sqrt{2\pi}\cdot\sigma_1} e^{-\frac{(X-a_1)^2}{2\sigma_1^2}} + p_2 \frac{1}{\sqrt{2\pi}\cdot\sigma_2} e^{-\frac{(X-a_2)^2}{2\sigma_2^2}},$$

где $\Theta = (p_1, p_2, a_1, a_2, \sigma_1, \sigma_2)$ являются неизвестными параметрами ($p_1 + p_2 = 1$).

В этом случае функция правдоподобия запишется

$$\prod_{i=1}^{n} f(X_i) = L(p_1, p_2, a_1, a_2, \sigma_1, \sigma_2).$$

Рассмотрим поведение $f(X)$ как функции от Θ. Если $a_j \neq X_i$, то $f(X_i)$ является *ограниченной* функцией, так как

$$p_j \frac{1}{\sqrt{2\pi}\cdot\sigma_j} e^{-\frac{(X_i-a_j)^2}{2\sigma_j^2}} \leqslant \frac{1}{\sqrt{2\pi}\,|X_i - a_j|} e^{-1/2}$$

для любых p_j и σ_j. Если же $p_j > 0$ и $a_j = X_i$, то $f(X_i)$ стремится к бесконечности как $(1/\sigma_1)$ при $\sigma_1 \to 0$. Однако, учитывая конечность предела $f(X_l)$ при $l \neq i$, получаем, что при $a_1 = X_i$ и $\sigma \to \infty$ функция $L(p_1, p_2, a_1 = X_i, a_2, \sigma_1, \sigma_2)$ стремится к бесконечности, как $1/\sigma_1$ для любого $p_1 \neq 1$ и любых a_2 и σ_2, чего не происходит при $\sigma_2 = \sigma_1$, так как при $\sigma_2 = \sigma_1 = \sigma$

$$\lim_{\sigma \to 0} L(p_1, p_2, a_1 = X_i, a_2, \sigma, \sigma) = 0.$$

Таким образом, любой набор $p_1, p_2, a_1 = X_i, a_2, \sigma_1 = 0$, $\sigma_2 = 0$, $p_1 + p_2 = 1$ и $0 < p_1 < 1$ обращает в бесконечность функцию правдоподобия.

Обобщение примера на *многомерные* смеси нормальных классов не представляет труда. Для этого достаточно рассмотреть случай, когда компоненты наблюдений X_i какого-либо класса j линейно зависимы, т. е. $|\widehat{\Sigma}| \to 0$ при $\mathbf{a}_j = X_i$.

Пример показывает, что возможны ситуации, когда не выполняются условия сходимости итерационной процедуры ЕМ-алгоритма к оценкам максимального правдоподобия.

Оценивание числа компонентов (классов) в модели смеси распределений. До сих пор, описывая процедуру статистического оценивания неизвестных значений параметров в модели смеси, предполагали число k компонентов (классов) в правой части модели (6.6′) заданным. Однако в реальных задачах часто общее число искомых классов неизвестно, и, следовательно, параметр k приходится также оценивать по тем же исходным данным (6.7).

С этой целью воспользуемся тем, что для ряда последовательных значений $k = 1, 2, \ldots$ выше уже решены оптимизационные задачи вида (6.8), т. е. вычислены такие значения параметров $\widehat{\Theta}(k)$ ($k = 1, 2, \ldots$), при которых соответствующие логарифмические функции правдоподобия $\ln L(\Theta(k))$ достигают максимума, т. е. при каждом фиксированном значении k имеем

$$\ln L(\widehat{\Theta}(k)) = \sup_{\Theta} \ln L(\Theta(k)).$$

Воспользуемся известным асимптотическим результатом (см., например, [157, § 13.8]), в соответствии с которым статистика критерия отношения правдоподобия

$$2\,|\ln L(\widehat{\Theta}(k+1)) - \ln L(\widehat{\Theta}(k))| \tag{6.14}$$

при условии справедливости гипотезы H_k: «истинное число компонентов смеси равно k» и при некоторых условиях регулярности функции $L(\Theta)$ имеет распределение, сходящееся (при $n \to \infty$) к распределению χ^2 с числом степеней свободы, равным $q + 1$ (q — размерность параметра θ, от которого зависит функция, задающая компонент смеси, а $q + 1$ — разность размерностей параметров $\Theta(k+1)$ и $\Theta(k)$). Процедуру построения оценки $\widehat{k}$ для неизвестного числа классов k определим следующим образом: задавшись некоторой величиной α уровня значимости критерия, производим последовательную (по $k = 1, 2, \ldots$) проверку гипотезы H_k при альтернативе H_{k+1} с помощью статистики (6.14) (гипотеза H_k отвергается, если величина (6.14) оказывается бо́льшей $100\,\alpha$ — процентной точки χ^2-распределения с $(q+1)$-й степенью свободы); величину $\widehat{k}$, при которой гипо-

теза H_k впервые оказалась неотвергнутой, принимаем за оценку истинного числа классов.

В [119] приводится результат, в соответствии с которым построенная таким образом оценка $\widehat{k}(n)$ дает при постоянных значениях уровня значимости α несколько *завышенные* величины числа классов, а именно имеет распределение (при истинном числе классов k_0):

$$\lim_{n\to\infty} P\{\widehat{k}(n) < k_0\} = 0;$$

$$\lim_{n\to\infty} P\{\widehat{k}(n) = k_0 + m\} = (1-\alpha)\,\alpha^m,\ m = 0,\ 1,\ 2,\ldots$$

Нетрудно подсчитать асимптотическую (по $n \to \infty$) величину среднего значения оценки $\widehat{k}$:

$$\lim_{n\to\infty} \mathbf{E}\,\widehat{k}(n) = k_0 + \frac{\alpha}{1-\alpha}.$$

Поэтому, если несколько модифицировать вышеописанную процедуру, выбирая в качестве уровней значимости критериев проверки гипотез H_k последовательности $\alpha(n)$, члены которых зависят от объема классифицируемых выборок и стремятся к нулю при $n \to \infty$, то можно добиться асимптотической несмещенности и состоятельности оценок $\widehat{k}(n)$.

Другие полезные приемы подбора подходящих значений неизвестного числа классов k основаны на различных методах разведочного статистического анализа, в частности на предварительной визуализации классифицируемых многомерных данных, например, с помощью процедур целенаправленного проецирования (см. раздел IV).

6.4.2. Процедуры, базирующиеся на методе моментов. Речь идет о процедурах решения системы уравнений метода моментов [11, соотношения (8.25)] применительно к рассматриваемой в данной главе модели смеси распределений.

При составлении системы уравнений метода моментов реализуется следующая схема:

1) используя знание общего вида функции плотности (полигона вероятностей) смеси $f(X)$ (см. формулу (6.6″)), вычисляют, в терминах неизвестных параметров $p_1, \ldots, p_k, \theta_1, \ldots, \theta_k$, всевозможные теоретические моменты компонентов $x^{(l)}$ исследуемого многомерного признака $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$: первые моменты $m_1^{(l)} = \mathbf{E}x^{(l)}$, вторые моменты $m_2^{(l_1 l_2)} = \mathbf{E}\,(x^{(l_1)} \cdot x^{(l_2)})$ и т. д. в количестве, равном общей размерности оцениваемого параметра Θ (если размерность параметра θ, определяющего распределение *внутри* класса, равна q, то общая размерность оцениваемого пара-

метра Θ при заданном числе компонентов смеси k составит $k - 1 + kq$);

2) по выборке классифицируемых наблюдений (6.7) подсчитываются соответствующие выборочные моменты $\widehat{m}_1^{(l)}$, $\widehat{m}_2^{(l_1 l_2)}$ и т. д.; 3) составляется система вида

$$\begin{cases} m_1^{(l)}(\Theta) = \widehat{m}_1^{(l)} \\ m_2^{(l_1 l_2)}(\Theta) = \widehat{m}_2^{(l_1 l_2)} \\ \ldots\ldots\ldots\ldots, \; l, \; l_1, \; l_2, \ldots = 1, \; 2, \; \ldots, \; p, \end{cases} \tag{6.15}$$

где левые части уравнений суть известные функции от неизвестных значений параметров $\Theta = (p_1, p_2, \ldots, p_{k-1}; \theta_1, \theta_2, \ldots, \theta_k)$, а правые части уравнений — известные числа.

Дальнейшие усилия направлены на решение системы (6.15), которое в каждом конкретном случае (при конкретизации общего вида компонентов $f(X, \theta_j)$ в (6.6′)) имеет свои специфические вычислительные трудности.

«Узкими местами» данного подхода являются чрезмерная громоздкость (а подчас практическая невозможность) его вычислительной реализации в случае многомерных анализируемых распределений $f(X, \theta_j)$ и большого числа k смешиваемых классов, весьма скромное качество статистических свойств получаемых при этом оценок $\widehat{\Theta}$ (в частности, дисперсия оценок $\widehat{m}_s$ для $s \geqslant 2$, а соответственно и дисперсия получаемых решений $\widehat{\Theta}$ остается слишком большой даже при возрастании объема выборки n). В работах [312, 178, 205, 291] содержатся примеры использования этого подхода для решения задачи расщепления смеси распределений, предпринимаются попытки преодолеть отмеченные выше трудности.

Пример 6.6 Исследование весового распределения хлопкового волокна по длине [159]. При решении некоторых задач из области технологии текстильной промышленности и, в частности, в задачах о вытягивании, смешивании, расчетах прочности пряжи, оценки неровности полуфабрикатов и т. п. необходимо исследовать весовое распределение хлопка по длине волокна. Предпринимавшиеся ранее специалистами попытки описать это распределение с помощью кривых Гаусса, Шарлье, Пирсона, закона χ^2 *«работали» лишь как формальная аппроксимация данной (обрабатываемой) выборки волокон и теряли свою работоспособность* при переходе к другим выборкам, поскольку не отражали самого механизма образования анализируемого распределения.

Визуальный анализ эмпирических плотностей весового распределения хлопкового волокна по длине, построенных по различным выборкам, позволил выявить некоторые общие (присущие всем экспериментальным кривым) закономерности (см. пунктирную кривую на рис. 6.3): каждая кривая имеет в зоне коротких волокон (в диапазоне от 15,5 до 21,5) небольшое, но устойчиво выраженное «плато» (близкое

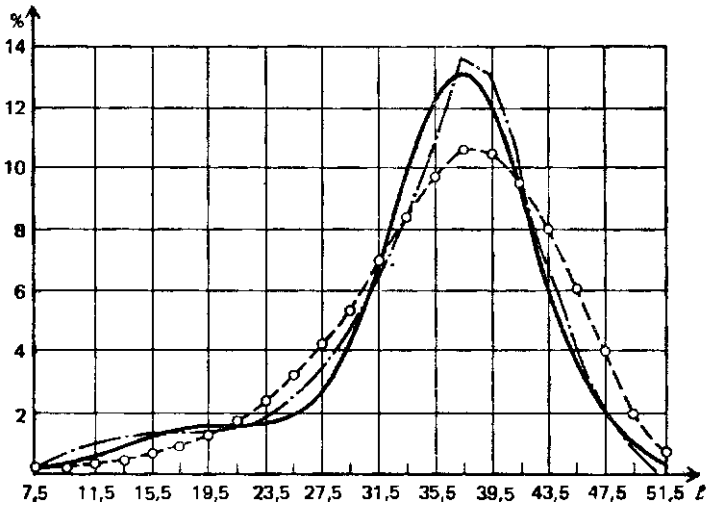


Рис. 6.3 Графики плотностей весового распределения хлопкового волокна по длине ($L_м$=38,1; $L_{ср}$=35,2; σ= =7,89) -·-·-·- — экспериментального, — O — O — — модельного χ^2, ——— — модельного, представленного смесью двух нормальных законов

к локальному максимуму) и, кроме того, четко выраженный глобальный максимум в диапазоне от 30 до 40,5 мм с формой кривой в этом диапазоне, близкой к нормальной. Это привело нас к гипотезе, что каждое из анализируемых распределений может быть представлено смесью двух нормальных распределений: первое из них (коротковолокнистое) $f_1(x)$ с относительно малым удельным весом p_1, небольшим средним значением a_1 и относительно большим коэффициентом вариации определяет закон распределения волокон в их короткой зоне, а второе (основное) $f_2(x)$ с преобладающим удельным весом $p_2 = 1-p_1$, средним значением $a_2 > a_1$ и относительно малым коэффициентом вариации $v_2 = \sigma_2/a_2$ определяет закон распределения волокон в их основной («длинной») зоне.

Итак, модель смеси (6.6′′) имеет здесь вид

$$f(x) = p_1 \varphi_1(x;\ a_1,\ \sigma_1^2) + p_2 \varphi_2(x;\ a_2,\ \sigma_2^2),$$

где

$$\varphi_j(x;\ a_j,\ \sigma_j^2) = \frac{1}{\sqrt{2\pi}\,\sigma_j} e^{-\frac{(x-a_j)^2}{2\sigma_j^2}}.$$

Поскольку нужно оценить пять неизвестных параметров p_1, a_1, σ_1^2, a_2, σ_2^2, то для построения системы уравнений метода моментов вида (6.15) необходимо, с одной стороны, вычислить в терминах этих параметров первые пять теоретических моментов исследуемой случайной величины $m_q(\Theta)$ $(q = 1, 2, \ldots, 5)$, а с другой — подсчитать те же самые моменты, но по имеющимся экспериментальным данным, т. е. вычислить выборочные моменты

$$\widehat{m_q} = \frac{\sum_{s=1}^{M} \omega_s \cdot x_s^q}{\sum_{s=1}^{M} \omega_s}.$$

Здесь M — число интервалов группирования по длине волокон; ω_s — вес волокон, отнесенных к s-му интервалу группирования; x_s — длина волокна, соответствующая середине s-го интервала группирования.

Переходя для удобства от моментов к семиинвариантам[1], получаем следующую систему уравнений относительно неизвестных p_1, p_2, a_1, a_2, σ_1^2 и σ_2^2:

$$\left.\begin{aligned}
&k_1 a_1 + k_2 a_2 = \lambda_1;\\
&k_1 k_2 (a_2 - a_1)^2 + k_1 \sigma_1^2 + k_2 \sigma_2^2 = \lambda_2;\\
&k_1 k_2 (k_1 - k_2)(a_2 - a_1)^3 + 3k_1 k_2 (a_2 - a_1)(\sigma_2^2 - \sigma_1^2) = \lambda_3;\\
&k_1 k_2 (k_1^2 - 4k_1 k_2 + k_2^2)(a_2 - a_1)^4 + 6k_1 k_2 (k_1 - k_2)(a_2 - \\
&- a_1)^2 (\sigma_2^2 - \sigma_1^2) + 3k_1 k_2 (\sigma_2^2 - \sigma_1^2)^2 = \lambda_4;\\
&k_1 k_2 (k_1^3 - 11k_1^2 k_2 + 11k_1 k_2^2 - k_2^3)(a_2 - a_1)^5 + 10k_1 k_2 (k_1^2 - \\
&- 4k_1 k_2 + k_2^2)(a_2 - a_1)^3 (\sigma_2^2 - \sigma_1^2) + 15k_1 k_2 (k_1 - k_2)(a_2 - \\
&- a_1)(\sigma_2^2 - \sigma_1^2)^2 = \lambda_5.
\end{aligned}\right\}$$

[1] Семиинварианты — некоторые вспомогательные характеристики распределения, определенным образом связанные с его моментами. В частности, для первых пяти семиинвариантов $\lambda_1, \ldots, \lambda_5$ имеют место следующие соотношения: $\lambda_1 = m_1$, $\lambda_2 = \mu_2$, $\lambda_3 = \mu_3$, $\lambda_4 = \mu_4 - 3\mu_2^2$ и $\lambda_5 = \mu_5 - 10\,\mu_2\mu_3$, где μ_j — q-й центральный момент анализируемой случайной величины.

В дальнейшем, правда, система была несколько модифицирована: в последнем пятом уравнении вместо λ_5 использовалась связь теоретических и экспериментальных *модальных* значений (x_{mod}).

При численном решении этой системы мы воспользовались методикой, номограммами и таблицами, предложенными в [312].

В табл. 6.1 приведены результаты 30-кратной численной «прогонки» этой системы: решалась задача расщепления 30 разных выборок.

Таблица 6.1

Номер выборки	x_{mod}	$\hat{m}_1$	$\hat{\mu}_3 = \hat{\sigma}_2^3$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{p}_2$	$\hat{a}_2$	$\hat{\sigma}_2$	$\hat{p}_1$	$\hat{a}_1$	$\hat{\sigma}_1$
1	25,0	23,5	24,70	— 80,3	2136,2	0,86	24,8	3,72	0,14	15,5	4,16
2	25,1	23,9	26,63	— 78,4	2248,7	0,82	25,2	4,05	0,18	17,9	5,36
3	26,0	24,6	30,47	—137,2	3428,2	0,86	26,0	4,07	0,14	16,1	5,62
4	26,2	24,7	25,30	— 87,2	2192,4	0,85	26,2	3,58	0,15	16,3	3,58
5	26,2	24,9	24,50	—110,7	2346,1	0,87	26,2	3,48	0,13	15,9	4,16
6	26,6	25,1	22,65	—107,5	2232,6	0,865	26,5	3,09	0,135	16,2	3,85
7	26,8	25,7	19,62	— 99,6	2088,7	0,82	27,0	2,87	0,18	19,7	5,28
8	26,8	25,5	27,55	—133,7	3191,6	0,81	27,0	3,66	0,19	19,0	6,0
9	27,2	25,6	26,52	—128,3	2921,2	0,82	27,5	3,50	0,18	19,5	6,75
10	27,7	25,9	26,73	—128,0	3012,8	0,80	27,5	3,54	0,20	19,6	5,81
11	30,5	28,2	39,44	—228,5	5864,1	0,86	30,55	4,0	0,14	16,9	4,84
12	30,6	28,2	33,90	—163,5	4420,0	0,84	30,1	3,86	0,16	18,3	4,20
13	30,7	26,6	49,10	—223,4	6740,0	0,70	30,5	4,19	0,30	18,5	4,68
14	30,8	28,3	40,45	—254,9	5760,8	0,856	30,2	4,19	0,144	17,0	5,29
15	31,9	29,9	39,19	—249,9	6592,5	0,88	31,5	4,31	0,12	18,0	5,44
16	31,9	29,1	41,90	—255,7	6390,5	0,83	31,4	4,03	0,17	18,0	4,45
17	31,9	29,6	42,60	—169,8	6012,2	0,75	32,0	5,06	0,25	23,5	6,21
18	32,3	29,0	44,10	—295,8	7248,5	0,79	31,5	3,96	0,21	19,5	6,11
19	32,5	30,5	67,40	—184,0	11829,2	0,70	33,0	6,81	0,30	25,0	8,41
20	32,6	30,3	45 70	—319,5	8375,1	0,85	32,5	4,15	0,15	17,8	4,86
21	36,6	32,6	56,80	—404,8	10923,7	0,82	35,1	4,86	0,18	21,2	6,84
22	36,6	33,6	50,69	—403,9	11539,0	0,89	35,34	4,84	0,11	19,4	6,62
23	36,7	33,3	56,25	—404,35	11639,8	0,85	35,7	4,79	0,15	19,8	5,52
24	37,6	34,2	63,68	—515,4	15699,4	0,86	36,6	5,14	0,14	19,3	5,95
25	37,6	34,4	51,80	—360,16	10383,7	0,86	36,5	4,62	0,14	21,2	6,14
26	38,0	34,9	78,10	—559,4	20922,5	0,83	37,5	6,20	0,17	22,0	8,20
27	38,1	35,2	62,25	—539,3	16307,7	0,87	37,5	5,16	0,13	20,0	6,18
28	38,2	34,2	66,26	—572,7	16833,5	0,83	36,9	5,07	0,17	21,2	7,76
29	38,9	35,9	69,89	—660,4	18780,5	0,85	38,4	5,35	0,15	21,9	6,07
30	40,5	38,3	59,00	—598,8	18148,3	0,90	40,2	5,17	0,10	21,9	7,76

Для всех 30 выборок независимо от селекционного сорта и модальной длины хлопкового волокна экспериментальные и теоретические (модельные) кривые плотностей графи-

чески хорошо совпадают как в центре диапазона, так и по краям. Более того, выведенная таким образом модель смеси распределений получила «задним числом» и *содержательное обоснование*, исходящее из механизма роста волокон хлопка. Данный пример показывает, как статистическое исследование может «натолкнуть» специалистов на некоторые содержательные выводы о физической природе изучаемого явления.

Построенная модель смеси позволила вывести важные новые и уточнить имевшиеся ранее соотношения между базовыми характеристиками распределения хлопкового волокна по длине, используемые в технологии текстильной промышленности.

6.4.3. Другие методы оценивания параметров смеси распределений. Практически каждую из существующих процедур статистического оценивания параметров смеси распределений можно отнести к одному из двух подходов. В первом из них (подход «от оценивания к классификации») исследователь начинает с решения задачи оценивания параметров смеси (6.6″), а затем переходит к собственно задаче классификации (если таковая стоит перед ним), причем решает ее, уже располагая оценками $\widehat{\theta}_j$ параметров θ_j каждого из компонентов смеси ($j = 1, 2, \ldots, k$), т. е., по существу, в рамках схемы ДА. К этому подходу относятся, в частности, представленные в п. 6.4.1 и 6.4.2 процедуры, базирующиеся на методе максимального правдоподобия и методе моментов. Практикуется также диаметрально противоположный по своей логической схеме подход (подход «от классификации к оцениванию»), при котором исследователь начинает с разбиения совокупности классифицируемых наблюдений на k подвыборок, а затем использует каждую (j-ю) из полученных подвыборок в качестве выборки из соответствующей (j-й) генеральной совокупности для оценивания ее параметров θ_j, после чего уточняет разбиение, и т. д.

К этому подходу можно отнести, в частности, описанный ниже *алгоритм адаптивного вероятностного обучения* (или «алгоритм SEM: *Stochastique — Estimation — Maximisation*» [202]), а также процедуры, базирующиеся на так называемом методе динамических сгущений [303, 302, 316, 106].

Алгоритм адаптивного вероятностного обучения (алгоритм SEM). Впервые предложен и проанализирован в [202]. По существу, авторы используют описанную в п. 6.4.1 схему ЕМ-алгоритмов, дополняя ее байесовской идеологией и этапом вероятностного обучения, которое реализуется в виде специальной процедуры генерирования на ЭВМ случайных

последовательностей. Прием вероятностного обучения с введением априорного распределения оцениваемых параметров использовался и ранее в задачах статистического оценивания, см., например, [314, 172, 310]. Использование алгоритма SEM позволяет в определенном (достаточно широком) классе идентифицируемых смесей решать (*в рамках основной процедуры*) задачу оценивания неизвестного числа k компонентов смеси (6.6'') и добиваться существенного снижения эффекта зависимости получаемого решения от исходной позиции начального приближения параметров алгоритма.

На исходной позиции алгоритма SEM *фиксируются*: начальное значение $\widehat{k}^{(0)} = \widehat{k}_{\max}$ для неизвестного числа компонентов смеси k (оно должно «с запасом» мажорировать истинное число классов k); некоторое пороговое значение $c(n, p)$, зависящее от объема n классифицируемой совокупности и от размерности p наблюдений X_i таким образом, что $c(n, p) \to 0$ при $n \to \infty$ (в [202] рекомендуется брать $c(n, p) = p/n$) и величины апостериорных вероятностей $g_{i1}^{(0)}$, $g_{i2}^{(0)}, \ldots, g_{i\widehat{k}_{\max}}^{(0)}$ $(g_{i1}^{(0)} + \ldots + g_{i\widehat{k}_{\max}}^{(0)} = 1)$ принадлежности случайно извлеченного из смеси наблюдения X_i соответственно к классу $1, 2, \ldots, \widehat{k}_{\max}$ $(i = 1, 2, \ldots, n)$.

Далее в следующей хронологической последовательности итерационного взаимодействия реализуются составляющие алгоритм SEM этапы «*Stochastique*» (*статистическое моделирование*), «*Maximisation*» (*максимизация функционала метода максимального правдоподобия*) и «Estimation» (*оценивание параметров смеси*).

Статистическое моделирование (v-я итерация, $v = 0, 1, 2, \ldots$). Последовательно для каждого $i = 1, 2, \ldots, n$ с помощью метода статистического моделирования Монте-Карло [11, § 3.3, 6.3] генерируются («разыгрываются») значения

$$e_j^{(v)}(X_i) = \begin{cases} 1 & \text{с вероятностью } g_{ij}^{(v)}, \\ 0 & \text{с вероятностью } 1 - g_{ij}^{(v)} \end{cases} \tag{6.16}$$

$(j = 1, 2, \ldots, \widehat{k}^{(v)};\ i = 1, 2, \ldots, n)$

полиномиально распределенных (с параметрами $g_{i1}^{(v)}$, $g_{i2}^{(v)}$, $\ldots, g_{i\widehat{k}^{(v)}}^{(v)}$) случайных величин $e_j^{(v)}(X_i)$. При этом производится по одному испытанию для каждой фиксированной пары (i, j).

Полученные реализации

$$e^{(v)}(X_i) = (e_1^{(v)}(X_i),\ e_2^{(v)}(X_i),\ \ldots,\ e_{\widehat{k}^{(v)}}^{(v)}(X_i)),\ i = 1, 2, \ldots, n,$$

определяют разбиение $S^{(\nu)} = (S_1^{(\nu)}, S_2^{(\nu)}, \ldots, S_{\hat{k}^{(\nu)}}^{(\nu)})$ анализируемой выборки $X_1, X_2, \ldots, X_n$ на классы по следующему правилу:

$$S_j^{(\nu)} = \{X_i : e_j^{(\nu)}(X_i) = 1\}, \quad j = 1, 2, \ldots, \hat{k}^{(\nu)}, \tag{6.17}$$

причем если в какой-либо класс попало в соответствии с этим правилом наблюдений меньше, чем $n\,c\,(n, p)$ (т. е. меньше, чем $n \cdot p/n - p$), то этот класс изымается (аннулируется) из нашего дальнейшего рассмотрения, а общее число классов соответственно уменьшается (переходят от оценки общего числа классов $\hat{k}^{(\nu)}$ к $\hat{k}^{(\nu+1)} = \hat{k}^{(\nu)} - l$, где l — число таких «малонаселенных» классов). При этом апостериорные вероятности пересчитываются по формулам

$$\tilde{g}_{ij}^{(\nu)} = \begin{cases} 0, \text{ если } j \text{ — «малонаселенный» класс;} \\ \dfrac{g_{ij}^{(\nu)}}{\sum\limits_{l \in J^{\nu}} g_{il}^{(\nu)}}, \text{ если } j \text{ — «достаточно представительный» класс,} \end{cases}$$

где J^{ν} — множество номеров «достаточно представительных» (на ν-й итерации) классов и оставшиеся «непристроенными» наблюдения (попавшие в «малонаселенные» классы) снова «разыгрываются» по правилу (6.16), (6.17) с вероятностями $\tilde{g}_{ij}^{(\nu)}$ и ими таким образом доукомплектовываются $\hat{k}^{(\nu+1)} = \hat{k}^{(\nu)} - l$ достаточно представительных классов.

Максимизация (ν-я итерация). По существу, этот этап так же, как и последующий, посвящен оцениванию параметров: его название обусловлено тем, что для реализации основной части процедуры оценивания приходится максимизировать (по искомому параметру) соответствующие функции правдоподобия.

На этом этапе (в рамках ν-й итерации) вычисляются оценки $\hat{\Theta}_j^{(\nu+1)} = (\hat{p}_j^{(\nu+1)}, \hat{\theta}_j^{(\nu+1)})$ параметров компонентов смеси (6.6′) по выборкам $\{X_i \in S_j^{(\nu)}\}$, $j = 1, 2, \ldots, \hat{k}^{(\nu+1)}$:

$$\hat{p}_j^{(\nu+1)} = \frac{1}{n} \sum_{i=1}^{n} e_j^{(\nu)}(X_i).$$

Оценки $\hat{\Theta}_j^{(\nu+1)}$ определяются как решения оптимизационных задач вида

$$\sum_{X_i \in S_j^{(\nu)}} \ln f(X_i; \Theta_j) \to \sup_{\Theta_j}, \quad j = 1, 2, \ldots, \hat{k}^{(\nu+1)}$$

(некоторые вспомогательные приемы в случае, когда уравнения метода максимального правдоподобия не дают решения, рассмотрены в [302, 212]).

Отметим, что если математические ожидания $\mathbf{a}_j = (a_j^{(1)}, \ldots, a_j^{(p)})'$ и (или) ковариационные матрицы $\Sigma_j = (\sigma_{qr}(j))_{q,r=1,2,\ldots,p}$ полностью определяют распределения внутри j-го класса, $j = 1,2, \ldots, k$ (случай смесей нормальных, пуассоновских, экспоненциальных и других распределений), то очередная итерация оценок максимального правдоподобия $\widehat{\mathbf{a}}_j^{(\nu+1)}$ и $\widehat{\Sigma}_j^{(\nu+1)}$ имеет вид

$$\widehat{\mathbf{a}}_j^{(\nu+1)} = \frac{\sum_{i=1}^{n} e_j^{(\nu)}(X_i)\, X_i}{\sum_{i=1}^{n} e_j^{(\nu)}(X_i)};$$

$$\widehat{\Sigma}_j^{(\nu+1)} = \frac{\sum_{i=1}^{n} e_j^{(\nu)}(X_i)\,(X_i - \widehat{\mathbf{a}}_j^{(\nu+1)})\,(X_i - \widehat{\mathbf{a}}_j^{(\nu+1)})'}{\sum_{i=1}^{n} e_j^{(\nu)}(X_i)}.$$

Оценивание (ν-итерация). Отправляясь от найденных на предыдущем этапе оценок

$$\widehat{\Theta}_j^{(\nu+1)} = \left(\widehat{p}_j^{(\nu+1)},\ \widehat{\theta}_j^{(\nu+1)}\right),\ j = 1, 2, \ldots, \widehat{k}^{(\nu+1)},$$

определяем значения оценок апостериорных вероятностей $g_{ij}^{(\nu+1)}$ по формулам

$$g_{ij}^{(\nu+1)} = \frac{\widehat{p}_j^{(\nu+1)} f\left(X_i;\ \widehat{\theta}_j^{(\nu+1)}\right)}{\sum_{l=1}^{\widehat{k}^{(\nu+1)}} \widehat{p}_l^{(\nu+1)} f\left(X_i;\ \widehat{\theta}_l^{(\nu+1)}\right)},$$

после чего переходим к следующей $(\nu + 1)$-й итерации этапа «статистическое моделирование».

Свойства алгоритма SEM исследованы в [202] аналитически в простейшем (представляющем лишь методический интерес) случае смеси, состоящей из двух полностью известных распределений $f_1(X)$ и $f_2(X)$, так что неизвестным параметром задачи является лишь единственное число p_1 — удельный вес первого компонента смеси (априорная вероятность принадлежности наблюдения, случайно извлеченного

из смеси, к классу 1). Правда, с помощью метода статистического моделирования авторы [202] рассмотрели большое число модельных примеров смеси [1] и пришли к выводу, что алгоритм SEM расщепления смеси распределений типа (6.6'') обладает следующими преимуществами в сравнении с другими алгоритмами: а) он работает относительно быстро и его результаты практически не зависят от «исходной точки»; б) позволяет избегать выхода на неустойчивые локальные максимумы анализируемой функции правдоподобия и, более того, дает, как правило, глобальный экстремум этой функции; в) получаемые при этом оценки параметров смеси являются асимптотически несмещенными; г) позволяет оценивать (в рамках самой процедуры) неизвестное число классов (компонентов смеси).

Пример 6.7. Расщепление смеси пятимерных нормальных распределений с помощью алгоритма SEM ([202], метод статистического моделирования). Зададимся в качестве компонентов смеси тремя ($k = 3$) пятимерными ($p = 5$) нормальными распределениями с удельными весами (априорными вероятностями) $p_1 = 0{,}5$, $p_2 = p_3 = 0{,}25$, с векторами средних значений

$$\mathbf{a}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_2 = \begin{pmatrix} 0 \\ 0 \\ 3 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 0 \end{pmatrix}$$

и с ковариационными матрицами

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{\Sigma}_2 = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix},$$

[1] Напомним схему рассмотрения модельных примеров с помощью метода статистического моделирования. Задаются полным описанием анализируемого закона распределения вероятностей $f(x)$ (в нашем случае — величинами k, p_j и функциями $f(X; \theta_j)$, $j = 1, 2, \ldots, k$); в соответствии с этим законом генерируют на ЭВМ выборку $X_1, X_2, \ldots, X_n$; затем, «забывая» знание закона $f(X)$, используют для обработки этой выборки анализируемый алгоритм, после чего сравнивают полученные результаты с тем, что было задано «на входе» (т. е. с $f(X)$).

$$\Sigma_3 = \begin{pmatrix} 16 & 0 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 16 \end{pmatrix}.$$

Генерируем на ЭВМ с помощью метода статистического моделирования [11, п. 6.3.3] выборку $X_1, X_2, \ldots, X_n$ из 400 наблюдений ($n = 400$), извлеченную из генеральной совокупности с плотностью распределения вероятностей

$$f(X) = 0{,}5\varphi(X;\ \mathbf{a}_1,\ \Sigma_1) + 0{,}25\varphi(X;\ \mathbf{a}_2,\ \Sigma_2) + \\ + 0{,}25\varphi(X;\ \mathbf{a}_3,\ \Sigma_3),$$

где $\varphi(X;\ \mathbf{a}_j,\ \Sigma_j)$ — плотность пятимерного нормального распределения с вектором средних значений $\mathbf{a}_j$ и ковариационной матрицей Σ_j («значения» $\mathbf{a}_j$ и Σ_j для $j = 1, 2, 3$ заданы выше).

Заметим, что оценка векторов средних $\mathbf{a}_j$ и ковариационных матриц Σ_j по той части сгенерированных наблюдений, которая принадлежит j-й генеральной совокупности, дает (в качестве эмпирических аналогов $\mathbf{a}_{j_э}$ и $\Sigma_{j_э}$ заданных теоретических значений $\mathbf{a}_j$ и Σ_j):

$$\mathbf{a}_{1_э} = \begin{pmatrix} 0{,}032 \\ 0{,}962 \\ 0{,}049 \\ 0{,}066 \\ -0{,}108 \end{pmatrix}, \quad \mathbf{a}_{2_э} = \begin{pmatrix} -0{,}016 \\ -0{,}122 \\ 2{,}923 \\ 0{,}012 \\ -0{,}175 \end{pmatrix}, \quad \mathbf{a}_{3_э} = \begin{pmatrix} 0{,}272 \\ -0{,}290 \\ 0{,}197 \\ 3{,}787 \\ 0{,}143 \end{pmatrix};$$

$$\Sigma_{1_э} = \begin{pmatrix} 0{,}827 & 0{,}020 & 0{,}127 & 0{,}087 & -0{,}069 \\ 0{,}020 & 1{,}779 & 0{,}085 & -0{,}082 & 0{,}113 \\ 0{,}127 & 0{,}085 & 1{,}136 & 0{,}102 & -0{,}087 \\ 0{,}087 & -0{,}082 & 0{,}102 & 0{,}970 & -0{,}084 \\ -0{,}069 & 0{,}113 & -0{,}087 & -0{,}084 & 0{,}948 \end{pmatrix};$$

$$\Sigma_{2_э} = \begin{pmatrix} 3{,}990 & 0{,}560 & -0{,}057 & 0{,}488 & -0{,}355 \\ 0{,}560 & 3{,}660 & 0{,}494 & 0{,}236 & -0{,}724 \\ -0{,}057 & 0{,}494 & 1{,}062 & 0{,}226 & -0{,}242 \\ 0{,}488 & 0{,}236 & 0{,}226 & 2{,}871 & -0{,}512 \\ -0{,}350 & -0{,}724 & -0{,}242 & -0{,}512 & 3{,}413 \end{pmatrix};$$

$$\Sigma_{3a} = \begin{pmatrix} 16{,}850 & 0{,}594 & 0{,}474 & 0{,}093 & 1{,}085 \\ 0{,}594 & 17{,}818 & 1{,}252 & -0{,}032 & 0{,}963 \\ 0{,}474 & 1{,}252 & 3{,}514 & 0{,}463 & -0{,}244 \\ 0{,}093 & -0{,}032 & 0{,}463 & 1{,}088 & -0{,}462 \\ 1{,}085 & 0{,}963 & -0{,}244 & -0{,}462 & 18{,}444 \end{pmatrix}.$$

Применение к *«перемешанной»* (сгенерированной на ЭВМ) выборке объема $n = 400$ алгоритма SEM (при $\widehat{k}^0 = 6$ и $g_{ij}^{(0)} = 1/\widehat{k}^{(0)} = 1/6$ для всех $i = 1, 2, \ldots, 400$) дает следующие оценки параметров смеси: $\widehat{k} = 3$; $\widehat{p}_1 = 0{,}508$; $\widehat{p}_2 = 0{,}245$; $\widehat{p}_3 = 0{,}247$;

$$\widehat{a}_1 = \begin{pmatrix} 0{,}035 \\ -0{,}886 \\ 0{,}887 \\ 0{,}085 \\ -0{,}157 \end{pmatrix}, \quad \widehat{a}_2 = \begin{pmatrix} 0{,}021 \\ 0{,}074 \\ 2{,}947 \\ -2{,}027 \\ -0{,}144 \end{pmatrix}, \quad \widehat{a}_3 = \begin{pmatrix} 0{,}232 \\ -0{,}361 \\ 0{,}153 \\ 3{,}822 \\ 0{,}214 \end{pmatrix};$$

$$\widehat{\Sigma}_1 = \begin{pmatrix} 0{,}772 & 0{,}060 & 0{,}145 & 0{,}012 & -0{,}032 \\ 0{,}060 & 1{,}802 & -0{,}028 & -0{,}115 & 0{,}206 \\ 0{,}145 & -0{,}028 & 1{,}212 & 0{,}128 & -0{,}154 \\ 0{,}012 & -0{,}115 & 0{,}128 & 0{,}950 & -0{,}074 \\ -0{,}032 & 0{,}206 & -0{,}124 & -0{,}074 & 0{,}869 \end{pmatrix};$$

$$\widehat{\Sigma}_2 = \begin{pmatrix} 4{,}205 & 0{,}582 & -0{,}021 & 0{,}720 & -0{,}491 \\ 0{,}582 & 3{,}997 & 0{,}280 & 0{,}491 & -0{,}551 \\ -0{,}021 & 0{,}280 & 1{,}125 & 0{,}363 & -0{,}538 \\ 0{,}720 & 0{,}491 & 0{,}363 & 2{,}831 & -0{,}609 \\ -0{,}491 & -0{,}551 & -0{,}538 & -0{,}609 & 3{,}368 \end{pmatrix};$$

$$\widehat{\Sigma}_3 = \begin{pmatrix} 16{,}946 & 0{,}506 & 0{,}306 & 0{,}171 & 0{,}158 \\ 0{,}506 & 17{,}807 & 1{,}068 & 0{,}033 & 0{,}741 \\ 0{,}306 & 1{,}068 & 3{,}329 & 0{,}549 & 0{,}100 \\ 0{,}171 & 0{,}033 & 0{,}549 & 1{,}044 & -0{,}655 \\ 1{,}158 & 0{,}741 & 0{,}100 & -0{,}655 & 18{,}792 \end{pmatrix}.$$

Наконец, классификация анализируемых 400 наблюдений, произведенная на последней итерации этапа «Статистическое моделирование» (доставляющей, кстати, наибольшее значение исследуемой функции правдоподобия), дает следующую картину «перекрестной» классификации в срав-

ненин с исходным (правильным) отнесением сгенерированных наблюденнй по составляющим генеральным совокупностям (табл. 6.2):

Таблица 62

Классы полученные с помощью алгоритма SEM	Исходная (правильная) классификация			Распределение наблюдений по классам SEM
	класс 1	класс 2	класс 3	
	200	100	100	
S_1	191	11	1	203
S_2	9	86	3	98
S_3	0	3	96	99

Таким образом, в данном примере доля неправильно расклассифицированных с помощью алгоритма SEM наблюдений составила 6,75% (27 наблюдений из 400). Этот результат так же, как и точность оценивания параметров смеси, можно признать вполне удовлетворительным.

6.5. Рекомендации по определению «исходных позиций» алгоритмов расщепления смесей распределений

Из предыдущего материала главы следует, что эффективность используемых для расщепления смесей алгоритмов (скорость их сходимости, опасность стабилизации итерационной процедуры алгоритма на стационарной точке функции правдоподобия, не дающей ее глобального экстремума, статистические свойства получаемых оценок) существенно зависит *от выбора исходной позиции алгоритма*, т. е. от конкретных начальных приближений для числа классов, априорных или апостериорных вероятностей и т. п., которые используют на нулевой итерации алгоритма.

Поэтому обычно настоятельно рекомендуется предпослать каждому из таких алгоритмов этап так называемого *разведочного статистического анализа* классифицируемых данных (техника разведочного статистического анализа описана в разделе IV). Он предназначен для предварительного «прощупывания» геометрической и вероятностной природы совокупности анализируемых данных и, в частности, позволяет формировать рабочие гипотезы о числе классов, типе вероятностного распределения внутри каждого из классов, величинах априорных вероятностей принадлежности наблюдения каждому из классов и т. п. Одним из основных прие-

мов такого типа анализа является проецирование анализируемых многомерных наблюдений на плоскость таким образом, чтобы максимально сохранить при этом интересующие исследователя специфические особенности рассматриваемой совокупности данных, например наличие и общее число четко выраженных «сгустков» (классов) или эффект концентрации данных этой совокупности вдоль некоторой гиперповерхности размерности меньшей, чем размерность исходного признакового пространства (такие процедуры носят название *методов целенаправленного проецирования*; см раздел IV). Производимый затем визуальный анализ спроецированных на плоскость исходных данных позволяет генерировать рабочие гипотезы по поводу требуемых начальных приближений алгоритмов.

ВЫВОДЫ

1. В задачах классификации без обучения одной из распространенных математических моделей, используемых при описании механизма генерирования классифицируемых данных, является *модель смеси вероятностных распределений*, когда каждый класс интерпретируется как параметрически заданная одномодальная генеральная совокупность (при неизвестном значении определяющего ее параметра), а классифицируемые наблюдения — как выборка из смеси таких генеральных совокупностей

2 *Решить задачу расщепления смеси распределений (в выборочном варианте)* — это значит по имеющейся выборке классифицируемых наблюдений, извлеченной из генеральной совокупности, являющейся смесью генеральных одномодальных совокупностей известного параметрического вида, построить статистические оценки для числа компонентов смеси, их удельных весов и параметров, их определяющих В *теоретическом варианте* задача расщепления смеси заключается в восстановлении компонентов смеси и смешивающей функции (удельных весов) по заданному распределению всей (т е смешанной) генеральной совокупности и называется *задачей идентификации компонентов смеси*. Эта задача не всегда имеет решение.

3. Базовая идея, лежащая в основе принятия решения, к какой из k анализируемых генеральных совокупностей следует отнести классифицируемое наблюдение, является одной и той же как для модели дискриминантного анализа (классификация при наличии обучения, см. гл 1—4), так и

для модели смеси: и в том и в другом случае *наблюдение приписывают к той генеральной совокупности (к тому компоненту смеси), в рамках которой (которого) оно выглядит наиболее правдоподобным*. Однако главное отличие схемы параметрического ДА от схемы автоматической классификации, построенной на модели смеси распределений, — в способе оценивания неизвестных параметров, от которых зависят функции, описывающие классы (в первом случае — по обучающим выборкам, а во втором неизмеримо сложнее — в рамках одного из методов оценки параметров смеси распределений).

4. *Основными «узкими местами» подхода, основанного на методе максимального правдоподобия статистического оценивания параметров смеси распределений*, являются (помимо необходимости «угадать» общий параметрический вид распределения, задающего каждый из классов) требование ограниченности анализируемой функции правдоподобия, высокая сложность и трудоемкость процесса вычислительной реализации соответствующих процедур и медленная сходимость порождаемых ими итерационных алгоритмов.

5. «*Узкими местами*» *подхода, основанного на методе моментов* статистического оценивания параметров смеси распределений, являются громоздкость его вычислительной реализации (особенно в случае высоких размерностей анализируемых распределений и большого числа смешиваемых классов) и относительно невысокое качество статистических свойств получаемых при этом оценок.

6. Статистический анализ смесей распределений проводится обычно в рамках одной из двух логических схем. В первой из них реализуется логика «*от оценивания параметров смеси к классификации*» (EM-алгоритмы, основанные на методе максимального правдоподобия, методе моментов и т. д.). Во второй, напротив, идут «*от классификации к оцениванию*», затем, имея оценки параметров распределений внутри классов, уточняют классификацию и т. д. (алгоритм SEM адаптивного вероятностного обучения).

7. Исследование свойств алгоритмов SEM, в которых схема EM-алгоритмов дополнена байесовской идеологией и этапом вероятностного обучения (реализованным в виде специальной процедуры генерирования на ЭВМ случайных последовательностей), показало *следующие их достоинства*: а) они работают относительно быстро и их результаты слабо зависят от выбора «начальных приближений»; б) SEM позволяют практически избегать выхода (в процессе итераций) на неустойчивые локальные максимумы анализируемой функции правдоподобия; в) позволяют в рамках самой процеду-

ры оценивать неизвестное число классов (компонентов смеси).

8. Поскольку эффективность используемых для расщепления смесей алгоритмов в большинстве случаев существенно зависит от выбора их «исходной позиции», целесообразно каждому из таких алгоритмов предпослать этап так называемого *разведочного статистического анализа* (см. раздел IV), который позволяет сформировать рабочие гипотезы о числе классов, типе вероятностного распределения внутри каждого из классов, величинах априорных вероятностей и т. п.

Глава 7. АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ, ОСНОВАННАЯ НА ОПИСАНИИ КЛАССОВ «ЯДРАМИ»

7.1. Эвристические алгоритмы

Рассмотрим сначала методы автоматической классификации (АК), непосредственно опирающиеся на постановку задачи выделения в многомерном пространстве компактных групп точек. Такие методы и отвечающие им алгоритмы называются *эвристическими*, так как само понятие «компактная группа (облако) точек» не поддается строгой формализации. В прикладных задачах автоматической классификации они стали применяться одними из первых и до сих пор сохраняют большое значение в разведочном анализе данных благодаря наглядности интерпретации полученных результатов и, как правило, простоте реализации. Для ряда эвристических процедур с развитием теории АК были найдены функционалы качества разбиения на группы и тем самым формализовано соответствующее им понятие «компактности».

7.1.1. Параллельные процедуры. Процедуры, в которых выборка вся сразу поступает на классификацию, называются параллельными.

Алгоритм *k* эталонов. Следуя [56], приведем типичный пример эвристического алгоритма, основная идея которого заключается в том, что совокупность объектов, находящихся на одинаковом расстоянии от каждого из k эталонов (ядер), образует компактную группу.

Пусть для классификации имеется выборка $O_1, \ldots, O_n$, причем i-й объект O_i характеризуется вектором признаков $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})$. Рассмотрим p-мерное признаковое про-

странство X^p вместе с функцией $d(X_i, X_j)$, задающей в X^p расстояние (либо степень близости).

Схема алгоритма

1. Выберем k эталонов $X_1^*, \ldots, X_k^*$ и порог d_0.

2. Поставим в соответствие объекту O_i код из k двоичных символов $\varepsilon_i = (\varepsilon_i^{(1)}, \ldots, \varepsilon_i^{(k)})$, где $\varepsilon_i^{(l)} = 1$, если $d(X_i, X_l) \leqslant d_0$ и $\varepsilon_i^{(l)} = 0$ в противном случае.

3. Разобьем выборку на классы, относя к одному классу объекты с одинаковым кодом.

В зависимости от порога d_0 и геометрии выборки число классов может варьироваться от 1 до 2^k. Анализируя полученное разбиение выборки на классы, исследователь может уточнить выбор эталонов и перейти к следующей итерации алгоритма. Если в качестве эталонов взять векторы признаков объектов из данной выборки, то на вход алгоритма достаточно подать матрицу взаимных попарных расстояний $\{\rho_{ij} = d(X_i, X_j)\}$. В [56] рекомендуется набор эталонов составлять из k случайно выбранных точек. Укажем наиболее важные модификации алгоритма, связанные со способом выбора эталонов:

а) эталоны строятся на основе представлений экспертов (или какой-либо другой априорной информации о типичных представителях классов);

б) эталоны берутся из векторов, соответствующих представителям заведомо разных классов, но не обязательно типичных в своем классе. При этом достоинством алгоритма является то, что число эталонов не обязано равняться числу классов;

в) эталоны могут пониматься в расширенном смысле, как ядра в методе динамических сгущений (см. п. 7.4).

Следующая модификация рассматриваемой процедуры связана с возможностью варьировать порог сходства.

Алгоритм взаимного поглощения. Рассмотрим процедуру автоматической классификации, целесообразность которой обосновывается в [58] правилом формирования сплоченных коллективов людей по принципу взаимного интереса и симпатии. Ключевым словом здесь является «взаимный», т. е. подразумевается, что отношение «объект O_i близок (интересен) объекту O_j» не симметрично и объекты O_i и O_j объединяются, если не только O_i близок к O_j, но и наоборот.

Схема алгоритма в предыдущих обозначениях алгоритма k эталонов:

1. Объекту O_i поставим в соответствие порог $d_i > 0$, называемый его радиусом влияния, т. е. объект O_j считается близким к O_i (находится в сфере влияния объекта O_i), если $d(X_i, X_j) \leqslant d_i$.

2. Объекту O_i поставим в соответствие код $\varepsilon_i = (\varepsilon_i^{(1)}, \ldots, \varepsilon_i^{(n)})$, где $\varepsilon_i^{(l)} = 1$, если $d(X_i, X_l) \leqslant d_l$, и $\varepsilon_i^l = 0$ — в противном случае.

3. Выделим в выборке классы, относя набор объектов $O_{i_1}, \ldots, O_{i_m}$, $m \leqslant n$, к одному классу, если у их кодов $\varepsilon_{i_1}, \ldots, \varepsilon_{i_m}$ все координаты с номерами $l = i_1, \ldots, i_m$ равны 1.

4. Выделим в выборке минимальное число классов, объединение которых дает всю выборку.

Ясно, что алгоритм дает в общем случае нечеткую классификацию выборки. Геометрически каждому объекту O_i в признаковом пространстве X^p отвечает шар ζ_i радиуса d_i с центром в точке X_i. Классу $\{O_{i_1}, \ldots, O_{i_m}\}$ отвечает пересечение $\bigcap_{l=1}^{m} \zeta_{i_l}$ шаров ζ_{i_l}, содержащее все центры $X_{i_1}, \ldots, X_{i_m}$ и называемое областью взаимного поглощения данного класса [58]. В ряде задач основной целью классификации является покрытие признакового пространства областями взаимного поглощения классов.

Ясно, что настройка рассматриваемого алгоритма на специфику решаемой задачи осуществляется выбором порогов d_i, $1 \leqslant i \leqslant n$. Приведем примеры такого выбора, взятые из [58]:

а) $d_i = \max_j d(X_i, X_j) - \delta$;

б) $d_i = \min_j d(X_i, X_j) + \delta$;

в) $d_i = \dfrac{\sum_{j=1}^{n} m_{ij} d(X_i, X_j)}{\sum_{j=1}^{n} m_{ij}}$.

Здесь δ — некоторая константа; $m_{ij} > 0$ — весовые множители. Они задаются либо эвристически, либо на этапе разведочного анализа служат управляющими параметрами.

7.1.2. Последовательные процедуры. Процедуры, в которых элементы выборки поступают на классификацию по одному или малыми порциями, называются *последовательными*.

Приведем простой последовательный алгоритм классификации, в основе которого лежит предположение, что пред-

ставители одного класса не могут быть удалены друг от друга более чем на заданную пороговую величину.

Пусть на классификацию объекты поступают последовательно, например по одному. Если исходная информация представлена в форме матрицы «объект — свойство», то параметрами алгоритма являются функция близости (расстояние) $d(O_i, O_j)$ между объектами и пороговое значение d_0; если исходная информация представлена матрицей взаимных расстояний, то единственным параметром является пороговое значение d_0.

Схема алгоритма

1. Первый объект O_1 объявляется ядром e_1 первого класса.

2. Пусть на m-м шаге выделено k классов с ядрами $e_1, \ldots, e_k$.

Для поступившего объекта O_m:
если $d(O_m, e_1) \leqslant d_0$, то O_m относим к первому классу;
если $d(O_m, e_{l-1}) > d_0$ и $d(O_m, e_l) \leqslant d_0$, $2 \leqslant l \leqslant k$, то O_m относим к l-му классу;
если $d(O_m, e_l) > d_0$, $1 \leqslant l \leqslant k$, то O_m объявляется ядром e_{k+1} нового $(k+1)$-го класса.

Если функция $d(O_i, O_j)$ удовлетворяет неравенству треугольника, как, например, когда $d(O_i, O_j)$ — метрика, то объекты, отнесенные алгоритмом к одному классу, удалены друг от друга не более чем на $2d_0$.

7.2. Алгоритмы, использующие понятие центра тяжести

При решении практических задач полезно иметь набор простых быстродействующих алгоритмов классификации для выработки первых представлений о структуре данных в признаковом пространстве. Таким алгоритмам посвящен этот параграф. Модификации алгоритмов, приведшие к ряду важных многопараметрических семейств их, ориентированных на проверку более сложных гипотез, возникающих уже в ходе исследования, описаны ниже в этой главе и в гл. 10.

Пусть исходная информация о классифицируемых объектах представлена матрицей «объект — свойство», столбцы которой задают точки p-мерного евклидова пространства.

7.2.1. Параллельные процедуры. Опишем один из наиболее известных алгоритмов, модификациями которого являются

многие важнейшие алгоритмы, приведенные далее (общая схема таких модификаций дана в гл. 10).

Алгоритм k-средних. Единственным управляющим параметром является число классов, на которые проводится разбиение $S = (S_1, \ldots, S_k)$ выборки $\mathbf{X}$. В результате получается несмещенное разбиение (см. п. 5.4.5) $S^* = (S_1^*, \ldots, S_k^*)$.

Схема алгоритма

1. Выберем начальное разбиение $S^0 = (S_1^0, \ldots, S_k^0)$, где

$$S_l^0 = \{X_{l1}^0, \ldots, X_{ln_l}^0\}, \quad \bigcup_1^k S_l^0 = \mathbf{X}, \quad S_l^0 \cap S_{l'}^0 = \varnothing, \quad l \neq l'.$$

2. Пусть построено m-е разбиение $S^m = (S_1^m, \ldots, S_k^m)$. Вычислим набор средних $\mathbf{e}^m = (e_1^m, \ldots, e_k^m)$, где $e_l^m = \frac{1}{n_l} \sum_{j=1}^{n_l} X_{lj}^m$.

3. Построим минимальное дистанционное разбиение, порождаемое набором $\mathbf{e}^m$ и возьмем его в качестве $S^{m+1} = (S_1^{m+1}, \ldots, S_k^{m+1})$, т. е. (ср. п. 5.4.5):

$$S_1^{m+1} = \{X \in \mathbf{X} : d(X, e_1^m) = \min_{1 \leqslant l \leqslant k} d(X, e_l^m)\}$$

. . . . . .

$$S_l^{m+1} = \left\{X \in \mathbf{X} \setminus \bigcup_1^{l-1} S_{l'}^{m+1} : d(X, e_l^m) = \right.$$

$$\left. = \min_{l \leqslant l' \leqslant k} d(X, e_{l'}^m)\right\}, \quad 2 \leqslant l \leqslant k,$$

где $d(X, e) = \|\mathbf{X} - e\|$ — расстояние в R^p.

4. Если $S^{m+1} \neq S^m$, то переходим к п. 2, заменив m на $m+1$, если $S^{m+1} = S^m$, то полагаем $S^m = S^*$ и заканчиваем работу алгоритма.

Введем расстояние $d(\mathbf{X}, \mathbf{e})$ от точки $X \in R^p$ до множества $\mathbf{e} = (e_1, \ldots, e_k)$, где $e_l \in R^p$ по формуле

$$d(X, \mathbf{e}) = \min_{1 \leqslant l \leqslant k} d(X, e_l).$$

Тогда можно рассмотреть *статистический разброс выборки* $\mathbf{X}$ *относительно множества* $\mathbf{e} = (e_1, \ldots, e_k)$:

$$F(\mathbf{X}; \mathbf{e}) = \sum_{X \in \mathbf{X}} d(X, \mathbf{e})^2.$$

Определим *статистический разброс разбиения* $S = (S_1, \ldots, S_k)$ *выборки* $\mathbf{X}$ как разброс этой выборки относительно множества $\mathbf{e}(S) = (e_1(S), \ldots, e_k(S))$, где $e_l(S)$ — средний вектор класса S_l, т. е. положим

$$F(S) = F(\mathbf{X};\ \mathbf{e}(S)).$$

Непосредственно из построения минимального дистанционного разбиения следует формула

$$F(S) = \sum_{l=1}^{k} \sum_{X \in S^l} \left\| X - e_l(S) \right\|^2. \tag{7.1}$$

Так как на последовательности разбиений S^0, S^1, ..., S^m, ..., которая строится в алгоритме k-средних, функционал $F(S)$ не возрастает, причем $F(S^m) = F(S^{m+1})$, только если $S^m = S^{m+1}$, то для любого начального разбиения S^0 алгоритм через конечное число шагов заканчивает работу (см. гл. 10).

Содержательно процедура алгоритма k-средних направлена на поиск разбиения S^* выборки $\mathbf{X}$ с минимальным разбросом.

В ряде случаев начальное разбиение S^0 задается как минимальное дистанционное разбиение, порождённое некоторым набором точек $\mathbf{e}^0 = (e_1^0, \ldots, e_k^0)$. Результат классификации зависит от выбора $\mathbf{e}^0$. Обычно для проверки устойчивости результата рекомендуется варьировать выбор $\mathbf{e}^0$. В тех случаях, когда из априорных соображений нельзя сразу выбрать число классов k, его находят либо перебором, либо вместо алгоритма k-средних используется алгоритм ИСОМАД (Isodata), в котором k является параметром, настраиваемым в ходе классификации (см. §. 7.4).

Алгоритм Форель [1]. Единственным управляющим параметром является порог r — радиус шаров, которыми покрывается выборка $\mathbf{X}$. Пусть $D_r(e) \subset R^p$ — шар радиуса r с центром в точке e. *Подвыборка* $\mathbf{X}^1 = \mathbf{X} \cap D_r(e)$ *называется несмещенной* в $D_r(e)$, если ее средний вектор совпадает с e. Классификация при помощи алгоритма Форель разбивается на несколько последовательных этапов. На первом в выборке $\mathbf{X}$ выделяется несмещенная подвыборка $\mathbf{X}_1$ в некотором $D_r(e_1)$, которая объявляется первым таксоном. На втором этапе та же процедура применяется к выборке $\mathbf{X} \setminus \mathbf{X}_1$. Таким образом, достаточно описать алгоритм только для первого этапа.

---

[1] Первоначальное название ФОРЭЛ — ФОРмальный ЭЛемент.

Схема алгоритма

1. Выберем начальное разбиение $S^0 = (S_1^0, S_2^0)$ выборки $\mathbf{X}$.

2. Пусть построено m-е разбиение $S^m = (S_1^m, S_2^m)$. Вычислим средний вектор e^m класса S_1^m.

3. Построим разбиение $S^{m+1} = (S_1^{m+1}, S_2^{m+1})$, где

$$S_1^{m+1} = \{X \in \mathbf{X} : d(X, e^m) \leqslant r\};\ S_2^{m+1} = \mathbf{X} \setminus S_1^{m+1}.$$

4. Если $S^{m+1} \neq S^m$, то переходим к п. 2, заменив m на $m+1$, если $S^{m+1} = S^m$, то полагаем $S^m = \mathbf{X}_1$ и заканчиваем работу первого этапа алгоритма.

Пополним пространство R^p «точкой» $*$, такой, что $d(X, *) = r$ для всех $X \in R^p$. Тогда статистический разброс выборки $\mathbf{X}$ относительно множества $\mathbf{e} = (e, *)$, где $e \in R^p$, запишется в виде

$$F(\mathbf{X}; \mathbf{e}) = \sum_{X \in \mathbf{X}} d(X, \mathbf{e})^2 = \sum_{X \in S_1} \|X - e\|^2 + r^2 |\overline{S}_1|, \qquad (7.2)$$

где $S_1 = \{X \in \mathbf{X} \cdot d(X, e) = \|X - e\| \leqslant r\}$, $|\overline{S}_1|$ — число элементов в множестве $\mathbf{X} \setminus S_1$. При помощи функционала (7.2) в гл. 10 показано, что последовательность разбиений $S^0, \ldots, S^m, \ldots$, которая строится на первом этапе алгоритма Форель, стабилизируется и алгоритм через конечное число шагов заканчивает работу.

Применение алгоритма Форель для ряда последовательных значений $r_\nu = r_0 - \nu\Delta$, где $\Delta = \frac{r_0}{N}$, $\nu = 1, 2, \ldots, N-1$, позволяет оценить наиболее предпочтительное число классов для данной выборки. При этом основанием для выбора числа классов может служить многократное повторение одного и того же числа классов для нескольких последовательных значений r_ν и его резкое возрастание для следующего шага по ν.

На основе алгоритма первого этапа Форели строится целое семейство алгоритмов, целью которых является разбиение выборки на заданное число классов, покрытие выборки $\mathbf{X}$ областями более сложной формы, чем шары, и т. п. Подробное описание этого семейства см. в [75]. Имеется модификация алгоритма первого этапа Форели, в которой порог r является параметром, настраиваемым в ходе поиска первого сгустка $\mathbf{X}_1$ (см. алгоритм Пульсар в п. 7.3.1).

7.2.2. **Последовательные процедуры.** Пусть $X_1, \ldots, X_n, \ldots$ — последовательность точек пространства R^p, описывающая последовательно поступающие на классификацию объ-

екты O_1, ..., O_n, ... Предполагается, что последовательность сколь угодно длинная. В тех случаях, когда число классифицируемых объектов конечно, применяется стандартный прием зацикливания конечной последовательности X_1, ..., X_n по правилу $X_{n+q} = X_q$, $q = 1, 2, \ldots$.

Положим $X(n) = \frac{1}{n} \sum_{i=1}^{n} X_i$. Тождество $X(n+1) = X(n) + \frac{1}{n+1}(X_{n+1} - X(n))$ позволяет параллельные процедуры, использующие понятие центра тяжести, модифицировать в последовательные процедуры. В качестве основного примера рассмотрим алгоритм k-средних.

Схема алгоритма

1. Выберем набор центров $\mathbf{e}^0 = (e_1^0, \ldots, e_k^0)$, где $e_l^0 \in R^r$, и набор весов $\Omega^0 = (\omega_1^0, \ldots, \omega_k^0)$, где $\omega_l^0 > 0$.

2. Пусть на m-м шаге построены набор центров $\mathbf{e}^m = (e_1^m, \ldots, e_k^m)$ и набор весов $\Omega^m = (\omega_1^m, \ldots, \omega_k^m)$. Для вновь поступившей точки X_m вычислим

$$d_m = \min_{1 \leqslant l \leqslant k} d(X_m, e_l^m)$$

и построим новые наборы $\mathbf{e}^{m+1}$ и Ω^{m+1}:

если $d(X_m, e_1^m) = d_m$, то положим

$$e_1^{m+1} = e_1^m + \frac{1}{\omega_1^m + 1}(X_m - e_1^m), \quad \omega_1^{m+1} = \omega_1^m + 1;$$

$$e_l^{m+1} = e_l^m, \quad \omega_l^{m+1} = \omega_l^m, \quad 1 < l \leqslant k;$$

если $d(X_m, e_{l-1}^m) > d_m$ и $d(X_m, e_l^m) = d_m$, где $1 < l \leqslant k$, то положим:

$$e_l^{m+1} = e_l^m + \frac{1}{\omega_l^m + 1}(X_m - e_l^m), \quad \omega_l^{m+1} = \omega_l^m + 1;$$

$$e_{l'}^{m+1} = e_{l'}^m, \quad \omega_{l'}^{m+1} = \omega_{l'}^m, \quad l' \neq l.$$

3. Критерием сходимости алгоритма k-средних является получение набора центров $\mathbf{e}^s = (e_1, \ldots, e_k)$, для которого минимальное дистанционное разбиение последовательности, порождаемой набором $\mathbf{e}^s$, является несмещенным. В связи с этим алгоритм останавливается, если в течение N шагов *подряд* практически не происходит пересчет центров. Например, если последовательность точек получена зацикливанием выборки объема n, то в качестве N достаточно взять

n. В общем случае число N задается априори или оценивается в ходе работы алгоритма исходя из модели всей совокупности точек

Пусть $e^r = (e_1, .., e_h)$ — набор центров, полученный на последнем шаге алгоритма. Итогом классификации является минимальное дистанционное разбиение последовательности, порождаемое набором e^r

Более подробный анализ этого алгоритма можно найти в [9] Модификация этого алгоритма, в которой число классов k является параметром, настраиваемым в ходе классификации, приведена в п 7 3.2.

7.3. Алгоритмы с управляющими параметрами, настраиваемыми в ходе классификации

7.3.1. Параллельные процедуры. Рассматриваемые ниже алгоритмы ИСОМАД и Пульсар являются модификациями соответственно алгоритмов k-средних параллельного типа и Форель

Алгоритм ИСОМАД (Isodata) [1]. Основной процедурой в этом алгоритме, как и в алгоритме k-средних, является минимальное дистанционное разбиение, порожденное набором центров. Число классов заранее не фиксируется, а определяется в ходе классификации Для этого используется ряд вспомогательных *эвристических* процедур, параметрами которых регулируются характеристики межклассовой и внутриклассовой структуры выборки на этапах классификации Конфигурация (схема) ИСОМАД не является фиксированной, ее развитие отражает богатый опыт практического применения этого алгоритма

Опишем наиболее распространенный вариант (ср. с [21, 156]).

Параметры, определяющие процедуру классификации:

k — предполагаемое число классов;

k_0 — начальное (разведочное) число классов;

θ_n — минимально допустимое число элементов в классе (функция от n, где n — число элементов во всей выборке);

θ_s — порог внутриклассового разброса;

θ_c — порог межклассового разброса;

Q — максимально допустимое количество пар центров классов, которые можно объединить;

I — допустимое число циклов итерации.

[1] Isodata — **Iterative Self-Organizing Data Analysis Techniques** (ИСОМАД — **Итеративный СамоОрганизующийся Метод Анализа Данных**).

Конкретные значения параметров задаются на основе априорной информации либо на этапе разведочного анализа выбираются из общих соображений, а затем корректируются от итерации к итерации.

Пусть на классификацию поступила выборка $\mathbf{X} = \{X_1, \ldots, X_n\}$, где $X_i \in R^p$. Выберем начальный набор центров $\mathbf{e}^0 = (e_1^0, \ldots, e_{k_n}^0)$.

Схема алгоритма

1. Выбираются значения параметров.

2. Строится минимальное дистанционное разбиение $S = (S_1, \ldots, S_{k_0})$ выборки $\mathbf{X}$, порожденное набором центров.

3. Пусть n_l — число элементов в классе S_l. Составляется $\tilde{S} = (\tilde{S}_1, \ldots, \tilde{S}_{k_m})$ из классов S_l разбиения S, у которых $n_l \geqslant \theta_n$, где k_m — полученное (текущее) число классов. $\tilde{S}$ присваивается обозначение $S = (S_1, \ldots, S_{k_m})$.

4. Вычисляется набор центров $\mathbf{e} = (e_1, \ldots, e_{k_m})$ из средних векторов классов, входящих в разбиение S.

5. Вычисляется вектор $D = (D_1, \ldots, D_{k_m})$, где

$$D_l = \frac{1}{n_l} \sum_{X \in S_l} \| X - e_l \|, \quad l = 1, \ldots, k_m.$$

6. Вычисляется

$$\overline{D} = \frac{1}{n} \sum_{l=1}^{k_m} n_l D_l.$$

7. а) Если текущий цикл итерации последний, то переход к 11; б) если $k_m \leqslant k/2$, то переход к 8; в) если текущий цикл итерации имеет четный порядковый номер или $k_m \geqslant 2k$, то переход к 11; в противном случае переход к 8.

8. Для каждого класса S_l вычисляется вектор $\sigma_l = (\sigma_l^1, \ldots, \sigma_l^p)$, где

$$\sigma_l^j = \sqrt{\frac{1}{n_l} \sum_{X \in S_l} (X^{(j)} - e_l^{(j)})^2}, \quad j = 1, \ldots, p, \quad l = 1, \ldots, k_m.$$

9. В каждом векторе σ_l отыскивается координата

$$\sigma_l^{j_l} = \max_{1 \leqslant j \leqslant p} \sigma_l^j, \quad 1 \leqslant l \leqslant k_m.$$

10. Если $\sigma_l^{j_l} > \theta_s$ для некоторого l, причем

а) $D_l > \bar{D}$ и $n_l > 2\ (\theta_n + 1)$

или

б) $k_m \leqslant k/2$,

то класс S_l с центром e_l *расщепляется* на два новых класса S_l^+, S_l^- с центрами e_l^+ и e_l^-, где соответственно $e_l^{\pm} = e_l \pm \varepsilon_l$, $\varepsilon_l = (\varepsilon_l^1, \ldots, \varepsilon_l^p)$ и $\varepsilon_l^j = 0$, если $j \neq j_l$, $\varepsilon_l^{j_l} = \gamma\sigma_l^{j_l}$, $0 < \gamma \leqslant 1$. Если расщепление класса на этом шаге происходит, то переход к 2 с набором центров $(e_1, \ldots, e_{l-1}, e_l^+, e_l^-, e_{l+1}, \ldots, e_{k_m})$, в противном случае переход к 11.

11. Вычисляется матрица (d_{ij}) взаимных расстояний между центрами классов $d_{ij} = \|e_i - e_j\|$.

12. Расстояния d_{ij}, где $i < j$, сравниваются с порогом θ_c. Пусть $d_{i_1 j_1} \leqslant d_{i_2 j_2} \leqslant \ldots \leqslant d_{i_{Q_1} j_{Q_1}}$ — упорядоченная последовательность тех из них, которые меньше θ_c. Вычеркнем из этой последовательности $d_{i_{Q_1} j_{Q_1}}$, если и только если в наборе $(i_1, i_2, \ldots, i_{Q_1-1})$ встречается индекс i_{Q_1} либо в наборе $(j_1, j_2, \ldots, j_{Q_1-1})$ встречается индекс j_{Q_1}. Проделаем аналогичную операцию с $d_{i_{Q_1-1} j_{Q_1-1}}$ и так далее до $d_{i_2 j_2}$.

Пусть $d_{l_1 t_1} \leqslant d_{l_2 t_2} \leqslant \ldots \leqslant d_{i_{Q_2} j_{Q_2}}$ — полученная в результате последовательность. Заметим, что по построению $i_1 = l_1$, $j_1 = t_1$. Положим $q = \min(Q, Q_2)$.

13. Слияние классов. Для каждой пары (l_i, t_i), $1 \leqslant i \leqslant q$ классы S_{l_i} и S_{t_i} сливаются в класс $S_{l_i} \cup S_{t_i}$. Непосредственно из 12 следует, что, если на предыдущем шаге было k_m классов, то теперь остается $k_m - q$ классов совокупности, которым переиндексацией присваивается обозначение $S = (S_1, \ldots, S_{k_m - q_n})$. Вычисляется набор центров $e = (e_1, \ldots, e_{k_m - q})$ средних векторов классов, входящих в S.

14. Если текущий цикл итерации последний, то алгоритм заканчивает работу. В противном случае переход к 1, если пользователь решил изменить какой-либо из параметров алгоритма, либо переход к 2, если в очередном цикле итерации параметры не меняются. Завершением цикла итерации считается каждый переход к 1 либо к 2.

Алгоритм Пульсар. Этот алгоритм, как и алгоритм Форель, состоит из последовательности одинаковых этапов, на каждом из которых выделяется один компактный класс (сгусток). Но радиус шара (величина окна просмотра) не фиксируется, а меняется (пульсирует) в ходе классификации. Для этого в алгоритм включены управляющие параметры, позволяющие поиск окончательного радиуса

реализовать в виде процедуры стохастической аппроксимации.

Опишем этап выделения одного сгустка [42].

Параметры, определяющие процедуру классификации:

$r_{\min}$, $r^{\max}$ — минимальный и максимальный радиусы;

$n_{\min}$, $n^{\max}$ — минимальное и максимальное число элементов в классе;

$\nu_{\text{доп}}$ — допустимое число колебаний радиуса. (Говорят, что произошло колебание радиуса, если $\Delta r_m \times \times \Delta r_{m+1} < 0$, где $\Delta r_m = r_m - r_{m-1}$, r_m — значение радиуса на m-м шаге);

δ — порог, регулирующий скорость изменения радиуса.

Схема алгоритма

1. Выберем начальный центр e^0 и значения параметров.

2. Для радиуса $r_0 = \frac{r_{\min} + r^{\max}}{2}$ построим класс $S^0 = \{X \in \mathbf{X}: \|X - e^0\| \leqslant r_0\}$, вычислим число элементов n_0 в классе S^0 и присвоим ν (числу колебаний радиуса) значение $\nu_0 = 0$.

3. Пусть на m-м шаге для центра e^m выбран радиус r_m, построен класс $S^m = \{X \in \mathbf{X}: \|X - e^m\| \leqslant r_m\}$, подсчитано число его элементов n_m и значение $\nu = \nu_m$.

Положим

$$e^{m+1} = \frac{1}{n_m} \sum_{X \in S^m} X;$$

$$r_{m+1} = \begin{cases} \min(r + \gamma\delta,\ r^{\max}), \text{ если } n_m \leqslant n_{\min}; \\ \max(r - \gamma\delta,\ r_{\min}), \text{ если } n_m > n^{\max}, \text{ причем } \nu_m < \nu_{\text{доп}} \\ \qquad\qquad \text{или } e^{m+1} \neq e^m; \\ r_m — \text{в остальных случаях.} \end{cases}$$

Здесь $\gamma = (1 + \nu_m)^{-1}$. Порог $\nu_{\text{доп}}$ учитывается при выборе радиуса r_{m+1} только тогда, когда $e_{m+1} = e_m$ и одновременно $n_m > n^{\max}$.

Далее положим

$\nu_1 = \nu_0 = 0$ и для $m \geqslant 1$

$$\nu_{m+1} = \begin{cases} \nu_m, & \Delta r_m \cdot \Delta r_{m+1} \geqslant 0; \\ \nu_m + 1, & \Delta r_m \cdot \Delta r_{m+1} < 0. \end{cases}$$

4. Если $e_{m+1} = e_m$, $r_{m+1} = r_m$, то алгоритм заканчивает работу, в противном случае переходим к 3, заменив m на $m + 1$.

7.3.2. Последовательные процедуры. В качестве основного примера, следуя [58], опишем вариант последовательного алгоритма k-средних (см. п. 7.2.2), в котором число классов не фиксировано, а меняется от итерации к итерации, настраиваясь под влиянием управляющих параметров на структуру выборки.

Параметры, определяющие процедуру классификации:

k_0 — начальное число классов;

φ — минимально допустимое расстояние между центрами разных классов;

ψ — максимально допустимое удаление элемента от центра его класса;

I — допустимое число циклов итерации.

Пусть на классификацию поступает последовательность точек $\{X_1, \ldots, X_n, \ldots\}$. Первые k_0 точек возьмем в качестве начального набора центров $\mathrm{e}^0 = (e_1^0, \ldots, e_{k_0}^0)$, присвоим центрам веса $\omega_l^0 = 1$.

Схема алгоритма

1. Выберем значения параметров φ и ψ.

2. Проведем огрубление центров e^0.

Расстояние между двумя ближайшими центрами сравнивается с φ. Если оно меньше φ, то эта пара центров заменяется их взвешенным центром, которому присваивается вес, равный сумме соответствующих двух весов. К полученному набору из $(k_0 - 1)$ центров опять применяется процедура огрубления и так далее до тех пор, пока расстояние между любыми двумя центрами будет не меньше, чем φ. Пусть в результате получается набор центров $\widetilde{\mathrm{e}}^0 = (\widetilde{e}_1^0, \ldots, \widetilde{e}_{\widetilde{k}_0}^0)$, $\widetilde{k}_0 \leqslant k_0$ с набором весов $\widetilde{\Omega} = (\widetilde{\omega}_1^0, \ldots, \widetilde{\omega}_{\widetilde{k}_0}^0)$.

3. Для вновь поступившей точки X_{k_0+1} вычислим минимальное расстояние до центра:

$$d_1 = \min_{1 \leqslant l \leqslant \widetilde{k}_0} d\left(X_{k_0+1}, \widetilde{e}_l^0\right).$$

Если $d_1 > \psi$, то X_{k_0+1} объявляется центром $\widetilde{e}_{\widetilde{k}_0+1}^0$ с весом $\widetilde{\omega}_{\widetilde{k}_0+1}^0 = 1$. Если $d_1 \leqslant \psi$, то самый близкий к X_{k_0+1} центр заменяется взвешенным центром этого старого центра и точки X_{k_0+1}. Вес нового центра считается равным сумме соответствующих весов. Все остальные центры и их веса не пере-

считываются. Полученные в итоге наборы центров и весов обозначаются через $E^1 = (e_1^1, \ldots, e_{k_1}^1)$ и $\Omega^1 = (\omega_1^1, \ldots, \omega_{k_1}^1)$. Заметим, что $\widetilde{k}_0 \leqslant k_1 \leqslant \widetilde{k}_0 + 1$.

4. Цикл итерации состоит из шагов п. 1—3. Если пользователь решил изменить значение параметров φ или ψ, то переход к п. 1, если φ и ψ не меняется, то переход к 2.

5. После прохождения I циклов полученный набор центров $e^I = (e_1^I, \ldots, e_{k_I}^I)$ объявляется набором эталонов классов и используется для классификации последующих наблюдений по методу минимального дистанционного разбиения.

Выбор значений параметров признается удачным (удовлетворительным), если получаемая в результате классификация оптимальна или с точки зрения экспертов, или в смысле принятых функционалов качества разбиения.

7.4. Алгоритмы метода динамических сгущений

Изложим, следуя в основном [106], один общий подход к статистической обработке данных, предложенный Э. Диде и его сотрудниками и названный «методом динамических сгущений» — МДС (Méthode des Nuées Dinamiques — MND). Этот подход хотя и формулируется в терминах общей задачи классификации, но, по существу, при соответствующем подборе управляющих параметров индуцирует разнообразные методы решения следующего широкого класса задач:

1) разбиение данной совокупности объектов или признаков на некоторое число (известное заранее или нет) однородных классов — собственно проблема автоматической классификации;

2) снижение размерности (числа анализируемых показателей) массива исходных данных, отбор наиболее информативных показателей;

3) статистический анализ предпочтений, задача их типологизации и агрегирования;

4) статистический анализ линейных моделей регрессионного типа;

5) оптимальная (в рамках решаемой конкретной задачи) оцифровка анализируемых переменных;

6) статистический анализ «двухвходовых» таблиц сопряженности.

Основная идея МДС, являющаяся далеким обобщением идеи метода k-средних (см. п. 7.2.1, 7.2.2), достаточно отчетливо выражена словами Э. Диде: «Среди всех разбиений на k классов следует найти разбиение, относительно каждого

класса которого заданное «ядро» оказалось бы наиболее представительным. Понятие ядра ... *имеет самый широкий смысл*: ядро класса (т. е. группы точек) может быть подгруппой точек, центром тяжести, осью, случайной переменной и т. д.» [106, с. 25] (курсив наш. — *В. Б.*).

7.4.1. Основные понятия и общая схема метода. Пусть $\mathbf{X} = \{X_1, \ldots, X_n\}$ — исследуемое множество объектов, каждый из которых характеризуется p-мерным вектором признаков, т. е. $X_l = (x_l^{(1)}, \ldots, x_l^{(p)})$.

Пространством k покрытий S^k называется множество, каждый элемент которого $S = (S_1, \ldots, S_k)$ представляет собой систему подмножеств (классов) элементов $\mathbf{X}$, удовлетворяющих заданной структуре классов. На практике встречаются обычно следующие типы структур классов: разбиения, покрытия, иерархии.

Пространством представителей L называется множество, каждый элемент которого *может* служить представителем (ядром) класса элементов $\mathbf{X}$. Выбирается мера сходства $D(X, l)$ между объектом $X \in \mathbf{X}$ и представителем $l \in L$.

Например:

а) $L = R^p$, тогда $l \in R^p$ и $D(X, l) = \sum_{i=1}^{p} (x^i - l^i)^2$ — квадрат стандартного евклидова расстояния;

б) $L = R^p \times M$, где M — некоторое семейство расстояний (см. § 5.2). Тогда $l = (e, d)$, $e \in R^p$, $d \in M$ и $D(X, l) = d(X, e)$. Семейство M обычно задается в параметрическом виде, например: 1) $M = \{\mathbf{M}\} : d(X, e) = (X - e)' \times \mathbf{M}(X - e)$ — семейство махаланобисского типа, где $\mathbf{M}$ — положительно определенная, симметрическая матрица;

2) $M = \{\lambda = (\lambda^1, \ldots, \lambda^p) \in R^p, \lambda^i > 0, \prod_{i=1}^{p} \lambda^i = 1\} : d(X, e) = \sum_{i=1}^{p} \lambda^i |x^i - e^i|$ — семейство «сити блок» (City Block) [1].

Пространством k представительств L^k для пространства покрытий S^k называется множество наборов $l = (l_1, \ldots, l_k)$, $l_i \in L$.

Для построения представительства l покрытия $S = (S_1, \ldots, S_k)$:

1) выбирается пространство представителей L и мера сходства $D(X, l)$;

2) выбирается *функция представительства* g, относя-

[1] В литературе иногда метрики этого семейства называются манхэттенскими.

щая к классу S_i представителя l_i, т. е. $g(S_i) = l_i$. Например, $g(S_i) = \arg\min \sum_{X \in S_i} D(X, l)$.

Для построения покрытия $S = (S_1, \ldots, S_k)$, отвечающего представительству $l = (l_1, \ldots, l_k)$:

1) выбирается пространство покрытий S^k;

2) выбирается *функция назначения* f, с помощью которой каждый объект X получает «назначение» в тот или иной класс, т. е. $f(l) = S$. Например, $f(l)$ — минимальное дистанционное разбиение выборки $\mathbf{X}$, порожденное набором $l = (l_1, \ldots, l_k)$ относительно меры сходства $D(X, l)$.

Метод динамических сгущений состоит из следующих частей (этапов):

1. Выбор пространства покрытий S^k.

2. Выбор пространства представителей L и меры сходства $D(X, l)$.

3. Выбор оптимизируемого критерия $W(S, l)$, позволяющего, используя $D(X, l)$, измерить «степень адекватности» между *всяким* покрытием $S \in S^k$ и *всяким* представительством этого покрытия $l \in L$. Критерий W строится обычно так, чтобы он принимал только неотрицательные значения.

4. Постановка задачи минимизации критерия W; выбор функции представительства g и функции назначения f, позволяющих решать эту задачу.

5. Построение алгоритма динамических сгущений (ADC), состоящего в последовательном итеративном использовании функций f и g, начиная с некоторого покрытия $S^{(0)} \in S^k$ или представительства $l^{(0)} \in L^k$.

6. Изучение свойств сходимости алгоритма динамических сгущений.

7.4.2. Алгоритмы классификации. Основная цель настоящего раздела — изложить подход МДС к построению алгоритмов автоматической классификации. Подробное описание, практические рекомендации и примеры применения этих алгоритмов содержатся в [106].

У всех рассматриваемых ниже алгоритмов одинаковыми являются:

пространство k покрытий S^k — множество всех разбиений $S = (S_1, \ldots, S_k)$ совокупности $\mathbf{X}$ на k непересекающихся классов;

вид оптимизируемого критерия $W(S, l)$:

$$W(S, l) = \sum_{i=1}^{k} D(S_i, l_i),$$

где $S = (S_1, \ldots, S_k)$, $l = (l_1, \ldots, l_k)$, $D(S_i, l_i) = \sum_{X \in S_i} D(X, l_i)$

— разброс i-го класса S_i относительно представителя (ядра) l_i и $D(X, l_i)$ — некоторая мера сходства между объектом X и представителем l_i;

способ построения функции назначения f. При выбранных ядрах классов $(l_1, \ldots, l_k)$ и мерах сходства $D(X, l_i)$ объект X относится к i-му классу по правилу минимального дистанционного разбиения:

$$f(l_1, \ldots, l_k) = (S_1, \ldots, S_k),$$

где

$$S_1 = \{X \in \mathbf{X} : D(X, l_1) = \min_{1 \leqslant i \leqslant k} D(X, l_i)\}$$

. . . . . .

$$S_j = \left\{X \in \mathbf{X} \setminus \bigcup_{i=1}^{j-1} S_i : D(X, l_j) = \min_{j \leqslant i < k} D(X, l_i)\right\}, \quad 2 \leqslant j \leqslant k;$$

способ построения функции представительства g. При выбранном пространстве представителей L задается пространство представительств L^k как подмножество наборов $(l_1, \ldots, l_k) \in L \times \ldots \times L = (L)^k$, выделяемое в $(L)^k$ некоторыми условиями. Тогда для заданного разбиения $S = (S_1, \ldots, S_k)$ его представительство $g(S)$ находится как решение задачи условной минимизации критерия $W(S, l)$, т. е.

$$g(S) = \arg \min_{l \in L^k \subset (L)^k} W(S, l).$$

В наиболее важном частном случае, когда $L^k = (L)^k$, представительство $g(S)$ находится как решение задачи безусловной минимизации на L:

$$g(S) = (\tilde{g}(S_1), \ldots, \tilde{g}(S_k)),$$

где

$$\tilde{g}(S_i) = \arg \min_{l \in L} D(S_i, l).$$

Схема алгоритма

1. Выберем начальное разбиение $S^0 = (S_1^0, \ldots, S_k^0)$.
2. Пусть построено m-е разбиение $S^m = (S_1^m, \ldots, S_k^m)$. Найдем m-е представительство (набор ядер) $l^m = g(S^m)$.
3. Найдем $(m+1)$-е разбиение $S^{m+1} = f(l^m)$.
4. Если $S^{m+1} \neq S^m$, то переходим к 2, заменив m на $m+1$. Если $S^{m+1} = S^m$, то полагаем $S^m = S^*$ и заканчиваем работу алгоритма.

Для получения конкретного варианта алгоритма классификации по приведенной схеме достаточно описать только следующие его компоненты:

а) пространство представителей L;

б) меру сходства $D(X, l)$, $X \in \mathbf{X}$, $l \in L$;

в) условия, выделяющие пространство представительств L^k в $(L)^k$.

В приведенных ниже алгоритмах будем использовать терминологию и, по возможности, обозначения из [106].

Пусть $\Pi(\mathbf{X})$ — множество подмножеств исследуемой совокупности объектов $\mathbf{X} = \{X_1, \ldots, X_n\}$. Для $P \in \Pi(\mathbf{X})$ обозначим через $|P|$ число элементов (мощность) множества P. Предположим, что на $\mathbf{X}$ задана положительная нормированная мера (распределение массы) μ, т. е. функция $\mu(X) > 0$, $X \in \mathbf{X}$, такая, что $\sum\limits_{X \in \mathbf{X}} \mu(X) = 1$.

Опишем сначала компоненты а, б, в алгоритмов, в которых представителями классов являются подмножества точек классифицируемой совокупности $\mathbf{X}$, т. е. $L \subseteq \Pi(\mathbf{X})$.

1. *Метод ядерного разбиения*

а) $L = \Pi(\mathbf{X})$;

б) $D(X, P) = \sum\limits_{Y \in P \subset \mathbf{X}} \mu(X)\mu(Y)\, d(X, Y)$,

где $d(X, Y)$ — некоторая мера близости между объектами из $\mathbf{X}$;

в) $L^k = (L)^k$.

2. *Метод, использующий ядра фиксированной мощности*

а) $L = L(m) = \{P \in \Pi(\mathbf{X}) : |P| = m\}$;

б) $D(X, P) = \sum\limits_{Y \in P \subset \mathbf{X}} \mu(X) \dfrac{\mu(Y)}{m}\, d(X, Y)$;

в) $L^k = (L(m))^k$.

3. *Метод, использующий ядра переменной мощности*

а) $L = L(\mu) = \{P \in \Pi(\mathbf{X}) : \mu(P) \leqslant \mu_0\}$,
где $\mu(P) = \sum\limits_{Y \in P} \mu(Y)$ и μ_0 — фиксированный уровень массы ядра;

б) $D(X, P) = \sum\limits_{Y \in P \subset \mathbf{X}} \mu(X) \dfrac{\mu(Y)}{\mu(P)}\, d(X, Y)$;

в) $L^k = (L(\mu))^k$.

Следующая группа алгоритмов использует в качестве ядер точки пространства признаков (формальные объекты).

Будем считать, что пространством признаков является R^p, т. е. $\mathbf{X} \subset R^p$.

1. *Метод центра тяжести*

а) $L = R^p$;

б) $D(X, l) = \mu(X)\, d(X, l)$,

где $d(X, l) = (X - l)'\, \mathbf{M}\, (X - l)$ — квадрат расстояния махаланобисского типа и $\mathbf{M}$ — *фиксированная* симметрическая положительно определенная матрица.

Разброс i-го класса S^i относительно ядра $l_i \in R^p$ в этом случае имеет вид

$$D(S^i, l_i) = \sum_{X \in S^i} \mu(X)\,(X - l_i)'\, \mathbf{M}\,(X - l_i);$$

в) $L^k = (R^p)^k$.

Когда $\mu(X) = 1/n$, где $n = |\mathbf{X}|$ и $\mathbf{M}$ — единичная матрица, то этот алгоритм представляет собой алгоритм k-средних параллельного типа.

2. *Метод адаптивных расстояний*

а) $L = R^p \times \mathrm{Dist}$, где Dist — некоторое семейство мер сходства в R^p;

б) $D(X, l) = d(X, e)$, где $l = (e, d)$, $e \in R^p$, $d \in \mathrm{Dist}$;

в) $L^k \subset (L)^k$.

В [106] исследуются отдельно варианты с квадратичными и неквадратичными расстояниями.

Квадратичные расстояния. В этом случае $\mathrm{Dist} = \{\mathbf{M}\}$: $d(X, e) = (X - e)'\, \mathbf{M}\, (X - e)$ — семейство махаланобисского типа.

1. *Независимый выбор меры сходства для каждого класса*, т. е. $L^k = (L)^k$, и представитель $(e_i, \mathbf{M}_i)$ i-го класса S_i находится как решение задачи:

$$(e_i^*, \mathbf{M}_i^*) = \arg \min_{\substack{e \in R^p \\ \mathbf{M} \in \mathrm{Dist}_1}} \sum_{X \in S_i} (X - e)'\, \mathbf{M}\, (X - e),$$

где Dist_1 — подмножество в Dist, составленное из матриц с определителем, равным 1.

Решая эту задачу, получаем (см. гл. 11) $e_i = \bar{X}(S_i)$ — центр класса S_i, $\mathbf{M}_i = |\hat{\Sigma}_i|^{1/p} \Sigma_i^{-1}$, где Σ_i — ковариационная матрица элементов i-го класса.

Теоретико-вероятностная модель, в рамках которой этот алгоритм оптимален, описана в п. 5.4.6. Каждое из наблюдений $X \in \mathbf{X} = \{X_1, \ldots, X_n\}$ извлекается из одной из k нормальных генеральных совокупностей $N(e_i, \Sigma_i)$, $i = 1, \ldots, k$. При этом e_i и Σ_i неизвестны.

Модели, в которой каждое наблюдение извлекается из одной из k нормальных генеральных совокупностей $N(e_i,$

Σ), $i = 1, \ldots, k$, с общей ковариационной матрицей Σ, где $e_1, \ldots, e_k$ и Σ неизвестны, соответствует алгоритм:

2. *Мера сходства, общая для всех классов*, т. е. $L^k \subset (L)^k$ и состоит из наборов $(e_1, \ldots, e_k, \mathbf{M})$.

В этом случае для заданного разбиения $S = (S_1, \ldots, S_k)$ представительство $g(S) = (e_1, \ldots, e_k, \mathbf{M})$ находится как решение задачи:

$$g(S) = \arg \min_{\substack{(e_1, \ldots, e_k, \mathbf{M}) \\ \mathbf{M} \in \mathrm{Dist}_1}} \sum_{i=1}^{k} \sum_{X \in S_i} (X - e_i)' \mathbf{M} (X - e_i),$$

которую *нельзя расщепить* на k задач поиска представителей каждого класса в отдельности. Решая эту задачу, получаем (см. гл. 11):

$e_i = \overline{X}(S_i)$ — центр класса S^i;

$\mathbf{M} = |\mathbf{W}|^{1/p} \mathbf{W}^{-1}$, где

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{X \in S_i} (X - \overline{X}(S_i))(X - \overline{X}(S))' = \sum_{i=1}^{k} \sum_i, \quad \sum_i — \text{ко-}$$

вариационная матрица i-го класса.

Неквадратичные расстояния. В качестве основного примера рассматривается $\mathrm{Dist} = \{\lambda = (\lambda^1, \ldots, \lambda^p), \ \lambda^i > 0, \ \prod_{j=1}^{p} \lambda^j = 1\}$, семейство «сити блок», которое приводит к методу, использующему медианную оценку центра класса.

Представитель $(e_i, d_i(X, e))$ i-го класса S_i, где e_i — центр класса, а $d_i(X, e) = \sum_{j=1}^{p} \lambda_i^j |x^j - e^j|$ — мера сходства, ассоциированная с этим классом, находится как решение задачи:

$$(e_i, d_i) = \arg \min_{(e, d)} \sum_{X \in S_i} d_i(X, e).$$

Так как

$$\sum_{X \in S_i} \sum_{j=1}^{p} \lambda^j |x^j - e^j| = \sum_{j=1}^{p} \lambda^j \sum_{X \in S_i} |x^j - e^j|,$$

получаем $e_i = (e_i^1, \ldots, e_i^p)$, где $e_i^j = \arg \min_{e^j} \sum_{X \in S_i} |x^j - e^j|$, т. е. e_i^j — медиана значений j-го признака по всем элементам i-го класса S_i. Положим $\alpha = (\alpha^1, \ldots, \alpha^p)$, где $\alpha^j = \sum_{X \in S_i} |x^j - e_i^j|$. Тогда

$\lambda_i^* = (\lambda_i^1, \ldots, \lambda_i^p)$,

$$\lambda_t' = \arg\min\left\{\Sigma \lambda^j \alpha^j : \lambda^j > 0, \quad \prod_{j=1}^{p} \lambda^j = 1\right\},$$

т. е.

$$\lambda_t^j = \left(\prod_{q=1}^{p} \alpha^q\right)^{1/p} \Big/ \alpha^j.$$

7.4.3. Автоматическая классификация неполных данных. На практике встречаются ситуации, когда исходная информация о классифицируемых объектах представлена матрицей «объект — свойство» с пропущенными значениями. Например, в социологических обследованиях некоторые индивидуумы могут отказаться ответить на те или иные вопросы, отдельные данные могут оказаться «стертыми» и т.п.

Опишем алгоритм МДС автоматической классификации совокупности объектов $O_1, \ldots, O_n$, характеризуемой неполной матрицей данных. Большим достоинством подхода МДС к этой задаче является то, что он не требует предварительного восстановления пропущенных значений и максимально использует специфику разбиения совокупности объектов на классы по принципу минимального дистанционного разбиения, порожденного набором ядер.

Выберем некоторое число (неважно какое) в качестве метки пропущенного значения. Поставим в соответствие объекту O_i, $i = 1, \ldots, n$, пару $(\mathbf{A}_i, X_i)$, где $\mathbf{A}_i$ — диагональная $(p \times p)$-матрица, а $X_i \in R^p$. Диагональный элемент a_i^{jj} матрицы $\mathbf{A}_i$ равен 1, если у O_i известно значение j-го признака, а $a_i^{jj} = 0$ в противном случае. Координата x_i^j вектора $X_i = (x_i^1, \ldots, x_i^p)$ равна значению j-го признака, если $a_i^{jj} = 1$ и равна метке в противном случае.

Введем в R^p евклидову метрику при помощи некоторой положительно определенной симметрической матрицы $\mathbf{M}$ ($\mathbf{M}$-метрику).

Квадратом *псевдорасстояния* от пары $(\mathbf{A}_i, X_i)$ до произвольной точки $e \in R^p$ называется

$$d_{\mathbf{M}}((\mathbf{A}_i, X_i), e) = (X_i - e)' \mathbf{A}_i' \mathbf{M} \mathbf{A}_i (X_i - e_i).$$

Непосредственно из определения следует, что значение псевдорасстояния $d_{\mathbf{M}}((\mathbf{A}_i, X_i), e)$ не зависит от выбранного значения метки, поэтому можно говорить о псевдорасстоянии $d_{\mathbf{M}}(O_i, e)$ от объекта O_i до точки $e \in R^p$.

Пусть $\mu(O_i)$ — некоторая весовая функция (положительная нормированная мера) на исследуемой совокупности объектов $\{O_1, \ldots, O_n\}$.

Выберем некоторый класс $S_i = (O_{i_1}, \ldots, O_{i_q})$. Выражение $\sum_{O_j \in S_i} \mu(O_j)\, d_{\mathbf{M}}(O_j, e)$ естественно назвать *псевдоразбросом* класса S_i относительно точки $e \in R^p$, а точку

$$e_* = \arg\min_{e \in R^p} \sum_{O_j \in S_i} \mu(O_j)(X_j - e)'\, \mathbf{A}_j \mathbf{M} \mathbf{A}_j (X_j - e)$$

— псевдоцентром тяжести класса S_i.

Пусть $\mathbf{M} = \mathbf{I}_p$ — единичная матрица и S_i совпадает со всей совокупностью объектов $\{O_1, \ldots, O_n\}$. Положим $\mu_i = \mu(O_i)$. Тогда псевдоцентр тяжести вычисляется по формуле:

$$e_* = (e_*^1, \ldots, e_*^p),$$

$$e_*^q = \sum_{i=1}^{n} \frac{\mu_i}{\sum_{j=1}^{n} \mu_j a_j^{qq}}\, a_i^{qq}\, x_i^q.$$

В общем случае нетрудно показать [106], что если каждый из p признаков наблюдается по крайней мере на одном из объектов класса S_i, то матрица

$$\mathbf{B}_i = \sum_{O_j \in S_i} \mu(O_j)\, \mathbf{A}_j \mathbf{M} \mathbf{A}_j$$

является положительно определенной и псевдоцентр тяжести e_{*i} класса S_i однозначно вычисляется по формуле:

$$e_{*i} = \mathbf{B}_i^{-1} \left(\sum_{O_j \in S_i} \mu(O_j)\, \mathbf{M} \mathbf{A}_j X_j \right).$$

Возвращаясь к общей схеме алгоритмов классифнкацни МДС, получаем, что если в качестве меры сходства взять псевдорасстояние, а в качестве центра класса — псевдоцентр, то можно непосредственно перенести на случай неполных данных алгоритмы метода центра тяжести и метода адаптивных квадратичных расстояний, изложенные в п. 7.4.2. При реализации этих алгоритмов необходимо только предусмотреть коррекцию на тех шагах алгоритма, когда встречается класс, для которого существует хотя бы один признак, ненаблюдаемый у всех элементов этого класса. Продемонстрируем такую коррекцию на примере *алгоритма k-средних параллельного типа для неполных данных.*

Поставим в соответствие исследуемой совокупности объектов $\{O_1, \ldots, O_n\}$ набор $\{(a_1, X_1), \ldots, (a_n, X_n)\}$, где $a_i = (a_i^1, \ldots, a_i^p)$ — вектор диагональных элементов матрицы $\mathbf{A}_i$ (см. выше). В R^p для простоты фиксируем стандартное

евклидово расстояние и будем считать, что точки имеют одинаковые веса. Тогда меру сходства между объектом O_i и точкой $e \in R^p$ можно записать в виде

$$D(O_i, e) = \frac{1}{n} \sum_{j=1}^{p} a_i^j \left(x_i^j - e^j \right)^2.$$

Схема алгоритма

1. Выберем начальный набор центров $(e_1^0, \ldots, e_k^0)$, $e_j^0 \in R^p$.

2. Пусть на m-м шаге построен набор центров $(e_1^m, \ldots, e_k^m)$. Построим минимальное дистанционное разбиение $S^m = (S_1^m, \ldots, S_k^m)$ совокупности объектов, используя псевдорасстояние $D(O_i, e)$.

3. Для каждого класса $S_i = \{O_{i_1}, \ldots, O_{i_{n_i}}\}$ вычислим вектор $b_i = (b_i^1, \ldots, b_i^p)$, где $b_i^j = \sum_{l=1}^{n_i} a_{i_l}^j$. Построим набор центров $(e_1^{m+1}, \ldots, e_k^{m+1})$, где

$$e_i^{m+1} = \left(e_i^{m+1,\,1}, \ldots, e_i^{m+1,\,p}\right);$$

$$e_i^{m+1,\,j} = \sum_{l=1}^{n_i} \frac{a_{i_l}^j x_{i_l}^j}{b_i^j}, \text{ если } b_i^j \neq 0;$$

$$e_i^{m+1,\,j} = e_i^{m,\,j}, \text{ если } b_i^j = 0.$$

4. Если $e_i^{m+1} \neq e_i^m$ хотя бы для одного i, то переходим к 2, заменив m на $m+1$, в противном случае заканчиваем работу алгоритма.

7.5. Алгоритмы метода размытых множеств

При анализе социально-экономических и биологических систем, в ряде задач технической и медицинской диагностики встречаются ситуации, когда вопрос не в том, принадлежит ли данный объект O_i, $1 \leqslant i \leqslant n$ классу S_j, $1 \leqslant l \leqslant k$, а в том, *до какой степени* O_i принадлежит S_j.

В связи с этим большое развитие получили методы нечеткой классификации, в основе которых лежит представление о классе как о размытом (нечетком) множестве объектов, для которых переход от *принадлежности* к данному классу к *непринадлежности* постепенен, а не скачкообразен.

7.5.1. Основные понятия, функционалы качества разбие-

ния, постановка задач. Пусть $\{O_1, \ldots, O_n\} = O$ — исследуемая совокупность объектов. *Размытое подмножество* объектов задается при помощи функции S, сопоставляющей с объектом O_i число $S(O_i)$, называемое *степенью принадлежности* объекта O_i этому подмножеству. Предполагается, что $0 \leqslant S(O_i) \leqslant 1$ для всех $i = 1, \ldots, n$. Ясно, что подмножество в обычном смысле задается функцией, принимающей значение 1 на элементах этого подмножества и 0 — на остальных элементах.

Размытые подмножества $S_1, \ldots, S_k$ совокупности O образуют разбиение на нечеткие классы, если $\sum_{l=1}^{k} S_l(O_i) = 1$ для всех i. В случае когда $\sum_{l=1}^{k} S_l(O_i) \geqslant 1$ для всех i, говорят, что размытые подмножества образуют *покрытие* нечеткими классами. Таким образом, разбиение $S = (S_1, \ldots, S_k)$ на нечеткие классы задает отображение

$$S : O \to R^k : S(O_i) = (S_1(O_i), \ldots, S_k(O_i)),$$

сопоставляющее с объектом O_i k-мерный вектор $S(O_i)$ его принадлежностей к классам этого разбиения.

Рассмотрим сначала случай, когда исходная информация о классифицируемых объектах представлена матрицей $\rho = (\rho_{ij})$ попарных взаимных расстояний (близостей) объектов. Тогда качество разбиения S оценивается тем, насколько соответствующее ему отображение $S : O \to R^k$ искажает «геометрическую» конфигурацию совокупности O, описываемую матрицей ρ. Например, [193]:

$$Q(S) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \left[\sum_{t=1}^{k} \sigma^2 (s_i^t - s_j^t)^2 - \rho_{ij}^2 \right]^2 \right\},$$

где $s_i^t = S_t(O_i)$ и σ — параметр.

Задача классификации — найти

$$S^* = \arg\min_{(S,\, \sigma)} \left\{ Q(S) : \sigma \in R^1,\ S = (S_1, \ldots, S_k),\ S_l = \right.$$

$$\left. = (s_1^l, \ldots, s_n^l),\ s_i^l \geqslant 0,\ \sum_{l=1}^{k} s_i^l = 1,\ i = 1, \ldots, n \right\}.$$

В такой постановке задача нечеткой классификации представляет собой вариант задачи *многомерного метрического шкалирования* (см. гл. 16), в котором экстремум функционала $Q(S)$ ищут на подмножестве отображений $S = (S_1,$

..., S_k), выделяемом условиями $\{s_i^l \geqslant 0, \sum_{l=1}^{k} s_i^l = 1,\ i = 1,\ldots, n\}$. Ясно, что функционалы качества k-мерного метрического шкалирования могут служить функционалами качества разбиения на k нечетких классов. Среди таких функционалов отметим (см. [66])

$$Q(S) = \sum_{i=1}^{n} \sum_{j=1}^{n} (\tilde{\sigma\rho}_{ij} - \rho_{ij})^2 \rho_{ij}^{a},$$

который является частным случаем функционалов $Q_1(Z)$ (формула (16.8)) и $Q_2(Z)$ (формула (16.8′)).

Здесь $\tilde{\rho}_{ij} = (\sum_{l=1}^{k} (s_i^l - s_j^l)^2)^{1/2}$, σ — параметр, $a = a_1$, если $\tilde{\rho}_{ij} \geqslant \rho_{ij}$, и $a = a_2$, если $\tilde{\rho}_{ij} \leqslant \rho_{ij}$.

Пусть теперь исходная информация представлена матрицей «объект — свойство», т. е. совокупность объектов можно отождествить с набором p-мерных точек $\mathbf{X} = \{X_1, \ldots, X_n\}$, $X_i = (x_i^1, \ldots, x_i^p)$.

Предположим, что $X_i \in R^p$. Опишем критерии качества разбиения, аналогичные критериям, использующим понятие усредненного внутриклассового разброса. Выберем монотонную функцию φ на отрезке $[0, 1]$, такую, что $\varphi(0) = 0$ и $\varphi(1) = 1$.

Для размытого подмножества S в $\mathbf{X}$ и точки $e \in R^p$ положим

$$D(S, e; \varphi) = \sum_{i=1}^{n} \varphi(s_i) \| X_i - e \|^2,$$

где $s_i = S(X_i)$. Для обычного подмножества S (когда $s_i = 0$ либо 1) выражение для $D(S, e; \varphi)$ не зависит от выбора φ и называется разбросом этого подмножества относительно точки e. По аналогии общее выражение для $D(S, e; \varphi)$ назовем *φ-взвешенным разбросом размытого множества S*. *Центр размытого множества* S определим, естественно, как решение задачи

$$e(S; \varphi) = \arg\min_{e \in R^p} D(S, e; \varphi).$$

Имеем

$$e(S; \varphi) = \sum_{i=1}^{n} \frac{\varphi(s_i)}{\sum_{j=1}^{n} \varphi(s_j)} X_i.$$

Внутриклассовый разброс размытого множества S определим как разброс этого множества относительно его центра, т. е. $D(S; \varphi) = D(S, e(S, \varphi); \varphi)$.

Теперь пусть $S = (S_1, \ldots, S_k)$ — некоторое разбиение на нечеткие классы. Положим

$$Q(S) = \sum_{t=1}^{k} D(S_t; \varphi_t).$$

В задаче классификации, отвечающей этому критерию, весовые функции $\varphi_1, \ldots, \varphi_k$ являются параметрами, которые можно фиксировать либо подбирать в ходе классификации.

Широкий класс критериев качества разбиения на нечеткие множества получается, если использовать подход метода динамических сгущений (см. п. 7.4). Выберем некоторую меру сходства $D(X, e)$. Тогда, как и выше, для монотонной функции $\varphi(s)$, $s \in [0, 1]$, $\varphi(0) = 0$, $\varphi(1) = 1$ вводится *разброс размытого подмножества* S относительно представителя l по формуле:

$$D(S, l; \varphi) = \sum_{i=1}^{n} \varphi(s_i) D(X_i, l).$$

Пусть теперь $S = (S_1, \ldots, S_k)$ — некоторое разбиение на нечеткие классы и $l = (l_1, \ldots, l_k) \in L^k \subset (L)^k$ — некоторое представительство (см. 7.4.1). Положим

$$W(S, l) = \sum_{t=1}^{k} D(S_t, l_t; \varphi_t).$$

Критерий качества разбиения, соответствующий $W(S, l)$ в МДС, имеет вид $Q(S) = W(S, g(S))$, где $g(S)$ — функция представительства. В рассматриваемом случае, при $L^k = (L)^k$, $g(S) = (l_1, \ldots, l_k)$, где

$$l_t = \arg\min_{l \in L} \Sigma \varphi_t(s_i^t) D(X_i, l), \quad s_i^t = S_t(X_i).$$

7.5.2. Алгоритмы нечеткой классификации. Специфику этих алгоритмов покажем на алгоритме k-средних [192, 193].

Пусть $\mathbf{X} = \{X_1, \ldots, X_n\}$, где $X_i \in R^p$.

Управляющие параметры:

k — число классов;

$\mathbf{M}$ — симметрическая положительно определенная матрица, задающая расстояние в R^p

$$\rho^2(X, Y) = (X - Y)' \mathbf{M} (X - Y)$$

(далее для простоты будем предполагать, что $\mathbf{M} = \mathbf{I}_p$ — единичная матрица);
α — число, $1 < \alpha < \infty$, определяющее весовую функцию $\varphi(s) = s^\alpha$ (см. п. 7.5.1).

Схема алгоритма

1. Выберем начальное разбиение $S^{(0)} = (S_1^{(0)}, \ldots, S_k^{(0)})$ на k нечетких классов, т. е. массив $s^{(0)}$ из n k-мерных векторов $s_1^{(0)}, \ldots, s_n^{(0)}$, $s_i^{(0)} = (s_i^{(01)}, \ldots, s_i^{(0k)})$, $s_i^{(0k)} \geqslant 0$, $\sum_{t=1}^{k} s_i^{(0t)} = 1$ для всех $i = 1, \ldots, n$.

2. Пусть построено m-е разбиение $S^{(m)}$ в виде массива $s^{(m)}$ из n k-мерных векторов $s_1^{(m)}, \ldots, s_n^{(m)}$. Вычислим набор центров $e_1^{(m)}, \ldots, e_k^{(m)}$, где

$$e_t^{(m)} = \sum_{i=1}^{n} \frac{\left(s_i^{(mt)}\right)^\alpha}{\sum_{j=1}^{n} \left(s_j^{(mt)}\right)^\alpha} X_i.$$

3. Построим $(m+1)$-е разбиение $S^{(m+1)}$ в виде массива $\{s_1^{(m+1)}, \ldots, s_n^{(m+1)}\}$, порождаемое центрами $e_1^{(m)}, \ldots, e_k^{(m)}$, где

$$s_i^{(m+1)} = \arg\min \left\{ \sum_{t=1}^{k} (s^{(t)})^\alpha \| X_i - e_t^m \|^2 : s = (s^{(1)}, \ldots, s^{(k)}), \right.$$

$$\left. s^{(t)} \geqslant 0, \ \sum_{t=1}^{k} s^{(t)} = 1 \right\},$$

т. е. если $I_i^m = \{t \mid 1 \leqslant t \leqslant k : \| X_i - e_t^m \| = 0\}$, то

$$I_i^m = \varnothing, \ s_i^{(m+1)\,t} = \left(\sum_{t'=1}^{k} \left(\frac{\| X_i - e_t^m \|}{\| X_i - e_{t'}^m \|} \right)^{\frac{2}{\alpha - 1}} \right)^{-1};$$

$$I_i^m \neq \varnothing, \ s_i^{(m+1)\,t} = 0, \ t \notin I_i^m \text{ и } \sum_{t \in I_i^m} s_i^{(m+1)\,t} = 1.$$

4. Если $S_i^{(m+1)} \neq S_i^{(m)}$ для некоторого i, $1 \leqslant i \leqslant n$, то переходим к 2; если $S^{(m+1)} = S^{(m)}$, то полагаем $S^{(m)} = S^*$ и заканчиваем работу алгоритма.

При фиксированных k и $\mathbf{M}$ описаниый алгоритм k-средних представляет собой параметрическое семейство по α, где $1 < \alpha < \infty$. Построение нечеткого разбиения, порождаемого набором центров (см. 3), определено при $\alpha \to 1$ и $\alpha \to \infty$.

Пусть $\alpha \to 1$. Положим $\widehat{I}_t^m = \{t,\ 1 \leqslant t \leqslant k : \| X_t - e_t^m \| = \min_{t'} \| X_t - e_{t'}^m \|\}$. Тогда

$$s_t^{((m+1)\,t)} = \begin{cases} 0, & t \notin \widehat{I}_t^m, \\ \dfrac{1}{\left|\widehat{I}_t^m\right|}, & t \in \widehat{I}_t^m. \end{cases}$$

С другой стороны, непосредственно из определения нечеткого класса $s_t^{(m+1)}$ следует, что при $\alpha = 1$ он имеет вид:

$$s_t^{(m+1)} = \arg\min \left\{ \sum_{t=1}^{k} s^t \| X_t - e_t^m \|^2,\ s^{(t)} \geqslant 0,\ \sum_{t=1}^{k} s^{(t)} = 1 \right\}.$$

Решением этой задачи линейного программирования является любой k-мерный вектор, лежащий на грани симплекса $\{s \in R^k : s^{(t)} \geqslant 0,\ \sum_{t=1}^{k} s^{(t)} = 1\}$, выделяемого условиями $s^{(t)} = 0$, если $t \notin \widehat{I}_t^m$. В частности, любая из вершин этого симплекса, у которой все координаты, кроме одной с номером из множества $\widehat{I}_t^m$, равны 0. Таким образом получается, что минимальное дистанционное разбиение, порождаемое набором центров $e_1^m, \ldots, e_k^m$ — это одна из допустимых классификации в алгоритме Беждека при $\alpha = 1$.

Пусть $\alpha \to \infty$. Тогда $s_t^{((m+1)t)} = \frac{1}{k}$, т. е. классификация вырождается. Таким образом, в алгоритме Беждека брать слишком большие значения параметра α не имеет смысла.

Точно так же, как расписан алгоритм нечеткой классификации по методу k-средних, можно расписать соответствующие алгоритмы по всем методам, основанным на описании классов «ядрами», рассмотренным в § 7.2 и 7.4. Например, полностью сохраняется общая схема алгоритмов классификации по методу динамических сгущений (см. 7.4.2). В случае нечеткой классификации необходимо только использовать следующий способ построения функции назначения f при выбранных ядрах классов $(l_1, \ldots, l_k)$ и мерах сходства $D(X, l_t)$: с объектом X_t сопоставляется k-мерный

вектор $s_i^* = (s_i^{*1}, \ldots, s_i^{*k})$ принадлежности к классам, такой, что

$$s_i^* = \arg\min \left\{ \sum_{t=1}^{k} \varphi_t(s^t) D(X_i, l_t) : s = \right.$$

$$\left. = (s^{(1)}, \ldots, s^{(k)}), \ s^{(t)} \geqslant 0, \ \sum_{t=1}^{k} s^{(t)} = 1 \right\}.$$

Например, в качестве весовых функций $\varphi(s)$ можно, как и выше, взять функции s^α, $1 < \alpha < \infty$. Тогда такие алгоритмы нечеткой классификации по МДС при $\alpha \to 1$ перейдут в алгоритмы, вариантами которых являются алгоритмы, описанные в § 7.4.

Детальное описание алгоритмов нечеткой классификации и исследование их можно найти в [193].

7.6. Алгоритмы, основанные на методе просеивания (решета)

Общим для всего этого семейства алгоритмов является наличие следующих трех блоков [35]:

1) парное сравнение объектов $O_1, \ldots, O_n$ и составление для каждого объекта O_i кортежа $\{O_{i_1}, \ldots, O_{i_{ni}}\}$ «похожих» на него объектов;

2) упорядочение (оцифровка) выборки в соответствии с целью классификации;

3) классификация, моделирующая принцип решета Эратосфена [1]: на вход классификатора поступает выборка, упорядоченная в блоке 2. Первый объект объявляется типичным представителем первого таксона, в который включается кортеж похожих на него объектов, составленный в блоке 1, после чего этот кортеж удаляется (вычеркивается) из выборки. Типичным представителем следующего таксона объявляется первый из оставшихся объектов упорядоченной выборки, а кортеж похожих на него объектов включается во второй таксон, те же из них, которые не были удалены на предыдущем шаге, удаляются из выборки. Такая процедура классификации продолжается до тех пор, пока

[1] Одна из формулировок решета Эратосфена (III в. до н. э.): если в множестве натуральных чисел 2, 3, 4,.. зачеркнуть числа, кратные первым r *простым* числам 2, 3, 5 ..., p_r, то первое (наименьшее) незачеркнутое число будет простым

вся выборка или наперед заданная доля ее не будет раскласси-фицирована.

Классификация, проведенная алгоритмами семейства, в общем случае приводит к покрытию классами, а не разбиению, так как таксоны могут пересекаться. Это связано с тем, что хотя элементы выборки на этапе 3 последовательно исключаются из рассмотрения, тем не менее они оказывают существенное влияние на последующую классификацию, поскольку на этапе 2 они участвовали в упорядочении выборки и порядок этот после их исключения не пересматривается.

Переход внутри семейства от одного алгоритма к другому осуществляется настройкой управляющих параметров.

В блоке 1 такие параметры определяются видом входной информации. Так, если выборка характеризуется матрицей «объект — свойство», то оценка похожести происходит с помощью той или иной меры близости, которая тем самым становится параметром алгоритма. Если же входная информация представлена в виде матрицы попарных взаимных расстояний, то кортеж похожих объектов рассчитывается с использованием пороговых значений, которые либо задаются, либо оцениваются на основе анализа матрицы взаимосвязей.

В блоке 2 упорядочение элементов выборки, как правило, проводится при помощи функционала, который сопоставляет с каждым объектом O_i выборки число, характеризующее, насколько этот объект *типичен* для совокупности похожих на него $\{O_{i_1}, \ldots, O_{i_m}\}$ с точки зрения целей классификации. Для выбранного функционала $F(O_i) = F(O_i; O_{i_1}, \ldots, O_{i_m})$ полагаем $O_i < O_j$, если $F(O_i) > F(O_j)$; в случае $F(O_i) = F(O_j)$, если нет дополнительной информации, то полагаем $O_i < O_j$ при $i < j$. Блок 3 не содержит управляющих параметров, поэтому далее при описании конкретных алгоритмов классификации методом Эратосфена (АКМЭ) опускаем его.

Рассмотрим более детально алгоритмы семейства, реализующие выделение таксонов при помощи поиска локальных максимумов функции плотности распределения объектов в признаковом пространстве $\mathbf{X}$.

Алгоритм на основе модельной локальной плотности. Выберем в качестве модельной некоторую плотность распределения $f_0(X)$, имеющую единственный максимум в нуле, и пусть $\mathbf{X}_\lambda^p \subset \mathbf{X}^p$ — такая область, что $P_{f_0}\{X \in \mathbf{X}_\lambda^p\} = \lambda$.

Будем считать, что $\mathbf{X}_1$ является носителем плотности $f_0(X)$, т. е. $\mathbf{X}_1^p = \{X \in \mathbf{X}^p : f_0(X) > 0\}$. Возьмем

оценку плотности распределения $f(X)$ по выборке $\{X_1, \ldots, X_n\}$ в виде

$$\tilde{f}(X) = \sum_{i=1}^{n} \mu_i f_0(X - X_i), \quad \mu_i > 0, \quad \sum_{i=1}^{n} \mu_i = 1,$$

где $\{\mu_i\}$ — веса точек $\{X_i\}$. Тогда первые два блока алгоритма можно описать следующим образом:

1. Для каждого X_i выделим подвыборку $\mathbf{X}(i) = \{X_{i_1}, \ldots, X_{i_{m_i}}\}$, составленную из всех элементов X_j, для которых $X_i - X_j \in \mathbf{X}_\lambda$.

2. Упорядочим выборку $\{X_1, \ldots, X_n\}$ при помощи плотности $\tilde{f}(X)$, т. е. считая, что $X_i \leqslant X_j$, если $\tilde{f}(X_i) \geqslant \tilde{f}(X_j)$. Например, возьмем модельную локальную плотность в R^p вида

$$f_0(X) = f_{p,\alpha}(X) = c_{p,\alpha}\left(1 - \frac{\|X\|^2}{r^2}\right)_+^{\alpha - 1},$$

где $c_{p,\alpha} = \dfrac{\Gamma\left(\frac{p}{2} + \alpha\right)}{[(2\alpha + p)\pi]^{p/2}\Gamma(\alpha)\sigma^p}$, $r^2 = (2\alpha + p)\sigma^2$ и σ^2 — дисперсия случайного вектора в R^p с плотностью распределения $f_0(X)$ (см. § 20.2).

Рассмотрим семейство алгоритмов для области $\mathbf{X}_1^p = \{X \in R^p, \ \|X\| \leqslant r\}$. Это семейство имеет два управляющих параметра: r и α. Похожими на элемент X_i считаются все элементы выборки, отстоящие от него не более чем на r, т. е. $\mathbf{X}(i) = \{X_j : \|X_i - X_j\| \leqslant r\}$. При фиксированном r с ростом α в модельной плотности $f_{p,\alpha}(X)$ σ^2 убывает и в упорядочении выборки все большее значение начинает иметь разброс кортежа $\mathbf{X}(i)$ относительно X_i, т. е. на первые места выходят X_i с меньшим разбросом кортежа $\mathbf{X}_{(i)}$ относительно X_i.

Для модельной плотности $f_{p,2}(X)$ и одинаковых весов μ_i, $1 \leqslant i \leqslant n$, классификация проводится при помощи плотности распределения

$$\tilde{f}(X) = \frac{c_{p,2}}{n} \sum_{i=1}^{n} \left(1 - \frac{\|X - X_i\|^2}{r^2}\right)_+ =$$

$$= \frac{c_{p,2}}{nr^2} \sum_{\|X - X_i\| \leqslant r} (r^2 - \|X - X_i\|^2),$$

которая с точностью до константы $\frac{c_{p,2}}{nr^2}$ совпадает с функционалом качества в алгоритме Форель.

Ясно, что выделение первого таксона в алгоритме Форель представляет собой *градиентный* поиск локального максимума плотности $\widetilde{f}(X)$.

Таким образом, если выборка представляет собой объединение сгустков, каждый из которых полностью помещается в шар радиуса r, то алгоритм Форель даст практически тот же результат, что и алгоритм классификации методом Эратосфена (АКМЭ) для плотности $f_{p,2}(X)$. Следующий таксон алгоритмом Форель выделяется после того, как из выборки удалены точки первого таксона, и т. д., что в ряде случаев (например, не угадан порог r) приводит к сильной зависимости результата классификации от выбора начальной точки, в то время как в АКМЭ вообще отсутствует необходимость выбора начальной точки.

Алгоритм на основе ближайших соседей. В этом алгоритме классификация проводится, по существу, при помощи оценки плотности распределения, получаемой по методу «ближайших соседей».

Выберем параметр λ, $0 < \lambda < 1$, и возьмем в качестве m целую часть числа λn, где n — объем выборки, поступившей на классификацию. Первые два блока этого алгоритма будут следующими.

1. Для каждого X_i составим кортеж $\mathbf{X}(i) = \{X_{i_1} = X_i, \ldots, X_{i_{m+1}}\}$ из m ближайших к нему элементов.

2. Пусть $S(\mathbf{X}(i), X_i)$ — разброс кортежа $\mathbf{X}(i)$ относительно X_i. Упорядочим выборку $\{X_1, \ldots, X_n\}$, считая, что $X_i \leqslant X_j$, если $S(\mathbf{X}(i), X_i) \leqslant S(\mathbf{X}(j), X_j)$.

В блоке 2 можно использовать также следующий способ упорядочения выборки $\{X_1, \ldots, X_n\}$:

2′. Пусть $r_i(m)$ — расстояние от X_i до самого дальнего из m ближайших к нему соседей $\{X_{i_1} = X_i, \ldots, X_{i_{m+1}}\}$. Упорядочим выборку $\{X_1, \ldots, X_n\}$, считая $X_i \leqslant X_j$, если $r_i(m) \leqslant r_j(m)$.

Описанные здесь блоки 1 и 2′ вместе с блоком 3, общим для всех алгоритмов метода просеивания, составляют алгоритм Уишарта [332], предложенный им в качестве альтернативы агломеративной процедуры иерархической классификации по методу «ближайшего соседа» в тех случаях, когда она приводит к нехарактерным для исследуемого явления объединениям типа «цепочного эффекта».

ВЫВОДЫ

1. Описаны наиболее известные и хорошо зарекомендовавшие себя при решении прикладных задач алгоритмы разбиения исследуемой совокупности объектов на классы как при известном, так и при неизвестном заранее числе классов.
2. Общим для всех рассмотренных алгоритмов является то, что в них распределение объектов по классам (классификация) осуществляется при помощи сформированного в ходе классификации набора «ядер» классов. Понятие ядра класса в широком смысле обсуждается в п. 7.4.
3. Алгоритмы различаются правилами распределения объектов по классам, типом ядер, тем, является ли класс четким или нечетким (см. § 7.5) подмножеством исследуемой совокупности объектов, является ли набор управляющих параметров фиксированным или настраиваемым в ходе классификации (см. § 7.3), а также тем, как поступают объекты на классификацию: вся совокупность сразу (параллельные алгоритмы) или порциями по одному, по нескольку (последовательные алгоритмы).

Глава 8. ИЕРАРХИЧЕСКАЯ КЛАССИФИКАЦИЯ

8.1. Основные определения

Пусть $\mathbf{X} = \{X_1, \ldots, X_n\}$ — конечное множество.

Определение 8.1. Иерархией s на $\mathbf{X}$ называется система подмножеств (классов) $\{S : S \subset \mathbf{X}\}$, такая, что

1) $\mathbf{X} \in s$;

2) $\{X_i\} \in s$, $i = 1, \ldots, n$;

3) если классы S и S' из s имеют не пустое пересечение, то $S' \subset S$ либо $S \subset S'$.

Пример 8.1. Пусть $\mathbf{X} = \{X_1, \ldots, X_7\}$. Тогда система подмножеств $s = \{\{X_i\}, i = 1, \ldots, 7, \{X_1, X_2\}, \{X_3, X_4, X_5\}, \{X_1, X_2, X_6\}, \mathbf{X}\}$ является иерархией на $\mathbf{X}$.

Исследование структуры иерархий удобно вести в терминах теории графов [83].

Определение 8.2. Графом $G = G(s)$ иерархии s на $\mathbf{X}$ называется ориентированный граф (V, E), вершины $v \in V$ которого соответствуют множествам $S \in s$, а ребра $e \in E$ — парам (S', S), таким, что: $S' \neq S$, $S' \subset S$ и в s не существует $S'' \neq S'$, для которого $S' \subset S'' \subset S$.

Ребро $e = (S', S)$ изображается стрелкой с началом S' и концом S.

Пример 8.2. Граф $G = (V, E)$ иерархии s из примера 8.1 имеет множество вершин: $V = \{v_i = \{X_i\},\ i = 1, \ldots, 7,\ v_8 = \{X_1, X_2\},\ v_9 = \{X_3, X_4, X_5\},\ v_{10} = \{X_1, X_2, X_6\},\ v_{11} = \mathbf{X}\}$ (рис 8.1). В графе иерархии вершина может быть концом нескольких стрелок, но, как следует из 3) определения 8.1, она является началом только одной стрелки.

Определение 8.3. Иерархия называется бинарной, если любое множество $S \in s$, содержащее более одного элемента, является объединением множеств S' и S'' из s, где $S' \cap S'' = \varnothing$.

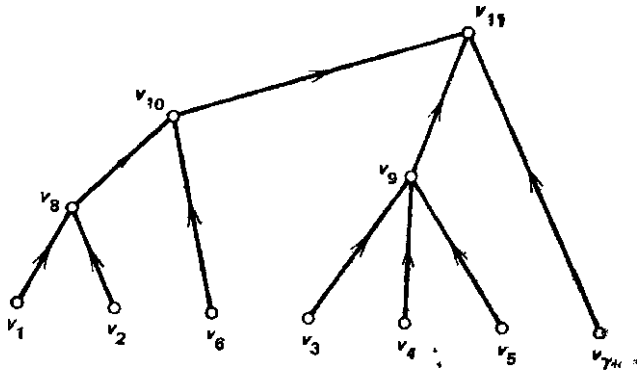


Рис. 8.1. Граф $G=(V, E)$ иерархии s из примера 8.1

Иерархия s из примера 8.1 не является бинарной, так как множество $\{X_3, X_4, X_5\}$ нельзя представить в виде объединения двух множеств из s. Нетрудно показать, что в любой бинарной иерархии s разбиение множества $S \in s$ на подмножества S' и S'' из s, т. е. представление $S = S' \cup S''$, однозначно. Иерархия является бинарной тогда и только тогда, когда в ее графе каждая вершина, соответствующая множеству, содержащему более одного элемента, является концом двух стрелок.

Определение 8.4. Иерархической классификацией данного множества объектов $\mathbf{X} = \{X_1, \ldots, X_n\}$ называется построение иерархии s на $\mathbf{X}$, отражающей наличие однородных, в определенном смысле, классов $\mathbf{X}$ и взаимосвязи между классами.

Алгоритмы иерархической классификации бывают: *дивизимные*, в которых множество $\mathbf{X}$ постепенно разделяется на все более мелкие подмножества, и *агломеративные*, в которых точки множества $\mathbf{X}$ постепенно объединяются во все более крупные подмножества.

Графы иерархий, полученных при помощи этих алгоритмов, называются соответственно дивизимными и агломеративными Если их изобразить на плоскости так, как на рис 8.2, то видно, что они описывают процедуру классификации при движении вверх по оси ординат Поэтому дивизимные алгоритмы называют также *нисходящими* (движение против стрелок), а агломеративные — *восходящими* (движение вдоль стрелок).

8.2. Методы и алгоритмы иерархической классификации

В основе алгоритмов иерархической классификации лежит тот или иной критерии качества $Q(S_1, \ldots, S_k)$ разбиения множества S на подмножества $S_1, \ldots, S_k$ (см. § 5.4). Обычно используются бинарные алгоритмы, когда $k = 2$ В этом случае $Q(S_1, S_2)$ имеет смысл близости $\rho(S_1, S_2)$ между множествами S_1 и S_2 (см § 5 3). Далее, говоря об алгоритмах иерархической классификации, будем иметь в виду только бинарные алгоритмы.

8.2.1. Дивизимные алгоритмы. Дивизимные алгоритмы строятся на принципе разделения множества S на подмножества (S_1^*, S_2^*), такие, что

$$(S_1^*, S_2^*) = \arg \max_{S_1 \cup S_2 = S} \rho(S_1, S_2). \qquad (8\ 1)$$

В реально используемых алгоритмах берется некоторое приближенное решение задачи (8.1), так как точное решение ее трудоемко даже при относительно небольшом объеме элементов в S Вид меры близости $\rho(S_1, S_2)$ может меняться в ходе алгоритма.

Изложим на важном примере основные приемы разделения класса S на подклассы S_1 и S_2 в дивизимных алгоритмах.

Пусть $\mathbf{X} = \{X_1, \ldots, X_n\}$, где $X_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \in R^p$ В качестве критерия однородности класса $S \subset \mathbf{X}$ возьмем статистический разброс

$$Q(S) = \sum_{X \in S} \|X - Z\|^2, \qquad (8.2)$$

где $Z = \frac{1}{|S|} \sum_{X \in S} X$ — центр класса S и $\|X - Z\|^2$ — квадрат евклидова расстояния между X и Z.

Положим

$$\rho(S_1, S_2) = Q(S_1 \cup S_2) - Q(S_1) - Q(S_2), \; S_1 \cap S_2 = \varnothing, \qquad (8.3)$$

т. е. мерой близости между классами считаем приращение статистического разброса при объединении классов.

Пусть $F(S_1, S_2) = Q(S_1) + Q(S_2)$. Для фиксированного класса S имеем: $\rho(S_1, S_2) + F(S_1, S_2) = \text{const}$. Следовательно, для решения задачи (8.1) разделения класса S достаточно найти

$$(S_1^*, S_2^*) = \arg \min_{S_1 \cup S_2 = S} F(S_1, S_2). \tag{8.4}$$

Функционал $F(S_1, S_2)$ является функционалом качества разбиения S на подклассы S_1 и S_2 в методе 2-средних, поэтому для получения классов S_1^* и S_2^* можно использовать алгоритмы 2-средних (см. п. 7.2.1).

Опишем теперь распространенный способ разделения множества S на подмножества при помощи линейного классификатора:

$$S_1 = \{X \in S : v' X \leqslant a\}, \; S_2 = \{X \in S : v' X > a\},$$

где $v \in R^p$, $\|v\| = 1$.

Для данного вектора $v \in R^p$, $\|v\| = 1$, рассмотрим проекцию v: $R^p \to R^1$: $X \to y = v'X$ и обозначим через $\widehat{S}$ образ множества S. Положим $\rho(S_1, S_2) = Q(\widehat{S}_1 \cup \widehat{S}_2) - Q(\widehat{S}_1) - Q(\widehat{S}_2)$, где $Q(\cdot)$ — как и выше, статистический разброс. Пусть

$$F_1(\widehat{S}_1, \widehat{S}_2) = \frac{1}{m_1}\left(\sum_{y \in \widehat{S}_1} y\right)^2 + \frac{1}{m_2}\left(\sum_{y \in \widehat{S}_2} y\right)^2.$$

Тогда имеет место формула:

$$\rho(S_1, S_2) = F_1(\widehat{S}_1, \widehat{S}_2) - \frac{1}{m}\left(\sum_{y \in S} y\right)^2,$$

где m, m_1 и m_2 — число элементов в множествах S, S_1 и S_2 соответственно. Следовательно, для фиксированного S достаточно найти:

$$(S_1^*, S_2^*) = \arg \max_{S_1 \cup S_2 = S} F_1(\widehat{S}_1, \widehat{S}_2). \tag{8.5}$$

Пусть $y_1 \geqslant y_2 \geqslant \ldots \geqslant y_m$ — упорядоченная числовая последовательность элементов множества $\widehat{S} \subset R^1$. Так как ищем на прямой точку a, разделяющую эту последовательность на две части, то

$$\widehat{S}_1 = \widehat{S}_1(m_1) = \{y_1, \ldots, y_{m_1}\}, \; \widehat{S}_2 = \widehat{S}_2(m_1) = \{y_{m_1+1}, \ldots, y_m\} \tag{8.6}$$

для некоторого m_1, т. е. в этом случае вместо метода 2-средних можно применить для решения задачи (8.5) метод последовательного перебора. Вычислив числовую последовательность $\{\varphi(m_1) = F_1(\widehat{S}_1(m_1), \widehat{S}_2(m_1)), m_1 = 1, \ldots, m-1\}$, получим пару $\widehat{S}_1^* = \widehat{S}_1(m_1^*)$, $\widehat{S}_2^* = \widehat{S}_2(m_1^*)$, где $m_1^* = \arg \max_{1 \leqslant m_1 \leqslant m-1} \varphi(m_1)$,

и, следовательно,

$$S_1^* = \{X \in S : v'X \leqslant y_{m_1^*}\},\ S_2^* = S \setminus S_1^*.$$

Таким образом описан алгоритм нахождения порогового значения a для линейного классификатора, задаваемого вектором $v \in R^p$, $\|v\| = 1$. В качестве v обычно берется какой-либо координатный вектор либо собственный вектор корреляционной матрицы множества $S \subset R^p$. Вектор v можно получить также как решение задачи целенаправленного проецирования (см. гл. 19), дающее проекцию $R^p \to R^1$, для которой $\widehat{S} = \{y_1, \ldots, y_m\}$ — наиболее неоднородная числовая выборка в смысле некоторого критерия.

Алгоритм иерархической классификации множества $\mathbf{X}$ из n элементов состоит из $(n-1)$ шагов. На вход дивизимного алгоритма подается все множество $\mathbf{X}$. На k-м шаге получается разбиение $S^{(k)}$ множества $\mathbf{X}$ на $(k+1)$ непересекающихся множеств (классов) $S_1^{(k)}, \ldots, S_{k+1}^{(k)}$, называемое разбиением k-го уровня.

Итоговая иерархия s представляет собой систему $s = \bigcup_{k=1}^{n+1} S^{(k)}$, образованную вложенными разбиениями $S^{(0)} \supset S^{(1)} \supset \ldots \supset S^{(n-1)}$. Здесь $S^{(0)} = \mathbf{X}$. (Говорят, что разбиение $S^{(1)} = (S_1^{(1)}, \ldots, S_{k_1}^{(1)})$ вложено в разбиение $S^{(2)} = (S_1^{(2)}, \ldots, S_{k_2}^{(2)})$, если каждый класс из $S^{(1)}$ является подклассом некоторого класса из $S^{(2)}$.) Если на k-м шаге получается разбиение $S^{(k)}$, все классы которого удовлетворяют выбранному критерию однородности, то алгоритм обычно останавливается.

8.2.2. **Агломеративные алгоритмы.** На вход агломеративного алгоритма подается разбиение $S^{(0)} = (S_1^{(0)}, \ldots, S_n^{(0)})$, где $S_i^{(0)} = \{X_i\}$. Разбиение k-го уровня имеет вид $S^{(k)} = (S_1^{(k)}, \ldots, S_{n-k}^{(k)})$ и строится из разбиения $S^{(k-1)}$, $k \geqslant 1$, путем объединения пары классов (S_1^*, S_2^*), где

$$(S_1^*, S_2^*) = \arg \min_{\substack{S_1 \neq S_2 \\ S_1, S_2 \in S^{(k-1)}}} \rho(S_1, S_2). \qquad (8.7)$$

Итоговую иерархию s образует система вложенных разбиений $S^{(0)} \subset S^{(1)} \subset \ldots \subset S^{(n-1)}$. Здесь $S^{(n-1)} = \mathbf{X}$.

Отметим, что иерархическая классификация при помощи бинарного алгоритма всегда дает бинарную иерархию.

Замечание. Иногда (см., например, [9]) иерархической классификацией множества $\mathbf{X}$ называют построение системы вложенных разбиений $S^{(0)} \supset S^{(1)} \supset \ldots \supset S^{(n-1)}$, где $S^{(0)} = \mathbf{X}$ или $S^{(0)} \subset S^{(1)} \subset \ldots \subset S^{(n-1)}$, где $S^{(n-1)} = \mathbf{X}$. Предыдущие рассуждения показывают эквивалентность такого определения иерархической классификации и данного выше определения 8.4.

Напомним, что наиболее употребительные в агломеративных алгоритмах меры близости $\rho(S_1, S_2)$ приведены в § 5.3. Свойства этих мер близости и методы эффективного решения задачи (8.7) обсуждаются ниже. Здесь же только отметим, что мера близости (8.3) удовлетворяет рекуррентному соотношению (5.10). Как будет видно далее, она обладает рядом важных свойств, которые обеспечивают широкое использование ее при решении задач классификации. В то же время ниже показано, что «Расстояние по центрам тяжести» (5.6) не обладает такими свойствами.

8.3. Графические представления результатов иерархической классификации

Алгоритмы иерархической классификации по сравнению, например, с алгоритмами, описанными в гл. 7, применимы для классификации множеств $\mathbf{X}$ относительно небольшого объема[1], но зато позволяют в ряде случаев получить более полный анализ структуры исследуемого множества объектов. Так, граф классификации дает представление о взаимосвязи между разбиениями $S^{(k_1)}$ и $S^{(k_2)}$ на разных уровнях, и если принято решение остановиться на разбиении $S^{(k)}$ миожества $\mathbf{X}$, то для каждого элемента $X \in \mathbf{X}$ можно оценить степень его принадлежпости к классу $S_l^{(k)}$ для некоторого l при помощи пути (простой цепи, см. [12, с. 147]), соединяющего вершины графа, соответствующие X и классу $S_l^{(k)}$.

Существенным преимуществом алгоритмов иерархической классификации является возможность наглядной интерпретации проведенного анализа. Имеется большая свобода

[1] При программной реализации алгоритмов иерархической классификации требования к объему оперативной памяти ЭВМ и время счета быстро растут с ростом числа элементов в множестве $\mathbf{X}$.

в построении на плоскости данного графа иерархии. Изучая различные изображения, согласованные с его структурой, можно получить ряд нетривиальных результатов об исследуемом множестве объектов, которые и рассмотрены ниже.

8.3.1. Индексации иерархии. Методы построения на плоскости графа иерархической классификации. Опишем общую схему построения на плоскости графа агломеративной классификации. Для получения изображения графа дивизимной классификации применимы те же методы, только во всех построениях надо поменять на противоположное направление оси ординат (см. рис. 8.2).

Определение 8.5 Индексацией ν иерархии называется отображение $\nu: s \to R^1$, ставящее в соответствие множеству $S \in s$ число $\nu(S)$ таким образом, что:

1) $\nu(S) = 0$ тогда и только тогда, когда S состоит из одного элемента;

2) $\nu(S') \leqslant \nu(S)$ для каждой пары (S', S) из s, такой, что $S' \subset S$.

Определение 8.6 Строгой индексацией ν иерархии s называется ее индексация, удовлетворяющая условию:

2') $\nu(S') < \nu(S)$ для каждой пары (S', S) из s, такой, что $S' \subset S$ и $S' \neq S$.

Пусть (s, ν) — некоторая индексированная агломеративная иерархия s на множестве $\mathbf{X} = \{X_1, \ldots, X_n\}$, построенная при помощи меры близости $\rho(S_1, S_2)$. Вершины графа этой иерархии, соответствующие множествам $\{X_i\}$, $i = 1, \ldots, n$, обозначим через v_i, а вершины, соответствующие множествам $S \in s$, $|S| > 1$, — через v_S. Допустим, что вершины v_{S_1} и v_{S_2} нанесены на плоскость, т. е. можно записать в координатах: $v_{S_1} = (v^1_{S_1}, v^2_{S_1})$ и $v_{S_2} = (v^1_{S_2}, v^2_{S_2})$. Тогда, если $S_1 \cup S_2 = S \in s$, то положим

$$v_S = \left(\frac{1}{2}(v^1_{S_1} + v^1_{S_2}),\ \nu(S)\right)$$

и соединим точки v_{S_1} и v_{S_2} с v_S. Далее будем использовать соединение, показанное на рис. 8.3. Начальный шаг построения обеспечивается тем, что, по предположению, вершины v_i располагаются на оси абсцисс. На этой оси они упорядочиваются так, чтобы в итоговом изображении графа минимизировать число пересечений ребер графа. Например, если ν — строгая индексация, то можно так упорядочить вершины v_i, $1 \leqslant i \leqslant n$, что ребра будут соединяться только в вершинах. Таким образом каждая индексация задает изображение графа.

До последнего времени в пакетах программ реализовывалось графическое представление агломеративной иерар-

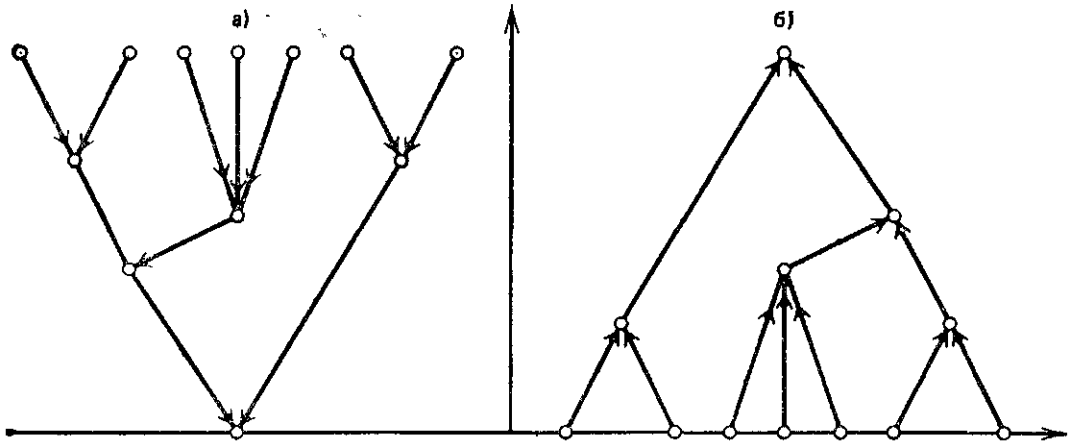


Рис. 8.2. Графы иерархий, описывающие процедуры классификации: *а*) дивизимная (нисходящая) процедура; *б*) агломеративная (восходящая) процедура

хии, основанное на строгой индексации ν, ставящей в соответствие множеству $S \in s$ номер шага, на котором это множество было включено в иерархию. В этом случае индексация ν любой иерархии s принимает значения $0, 1, \ldots, (n-1)$ [1]. Важные исследования по индексациям иерархий проведены в [248], на результаты [248] опирается изложение этих вопросов в книге.

8.3.2. Оцифровка изображения графа иерархической классификации. При наглядной интерпретации результатов

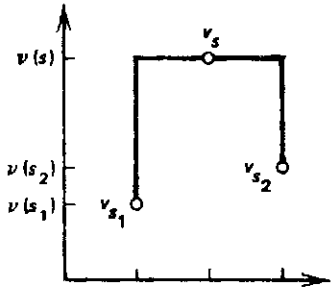


Рис. 8.3. Правило соединения вершин при изображении на плоскости графа иерархической классификации с использованием индексирующего отображения ν

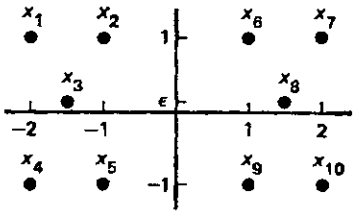


Рис. 8.4. Конфигурация точек множества $X = \{x_1, \ldots, x_{10}\}$ из примера 8.3

классификации большое значение имеет то, какую дополнительную информацию о структуре исследуемой совокупности объектов несет распределение на отрезке $[0,1]$ числовой последовательности $\{\nu(S)/\nu(\mathbf{X}),\ |S| > 1,\ S \in s\}$ для выбранной индексации ν.

Пример 8.3. Пусть $\mathbf{X} = \{X_1, \ldots, X_{10}\}$ — множество точек на плоскости, изображенное на рис. 8.4, и s — его агломеративная иерархия, построенная по мере близости (5.6): $\rho(S_1, S_2) = \|Z_1 - Z_2\|^2$, где Z_i — центр класса S_i, $i = 1,2$. Здесь $X_3 = (-1, 5, \varepsilon)$ и $X_8 = (1,5, \varepsilon)$, где $0 < \varepsilon < 0,125$. Поэтому

$$\|X_1 - X_3\|^2 = \|X_2 - X_3\|^2 > 1 + \varepsilon^2.$$

Из рис. 8.5 видна роль выбора индексирующего отображения ν. В случае а) наглядна последовательность вхождения классов S_i в иерархию; в случае б) это невозможно уви-

[1] На распечатках графа иерархии, изображенного с помощью такой индексации, обычно вместо номера уровня выводят значения меры близости $\rho(S_1, S_2)$ между классами S_1 и S_2, которые на этом уровне образовали новый класс $S = S_1 \cup S_2$.

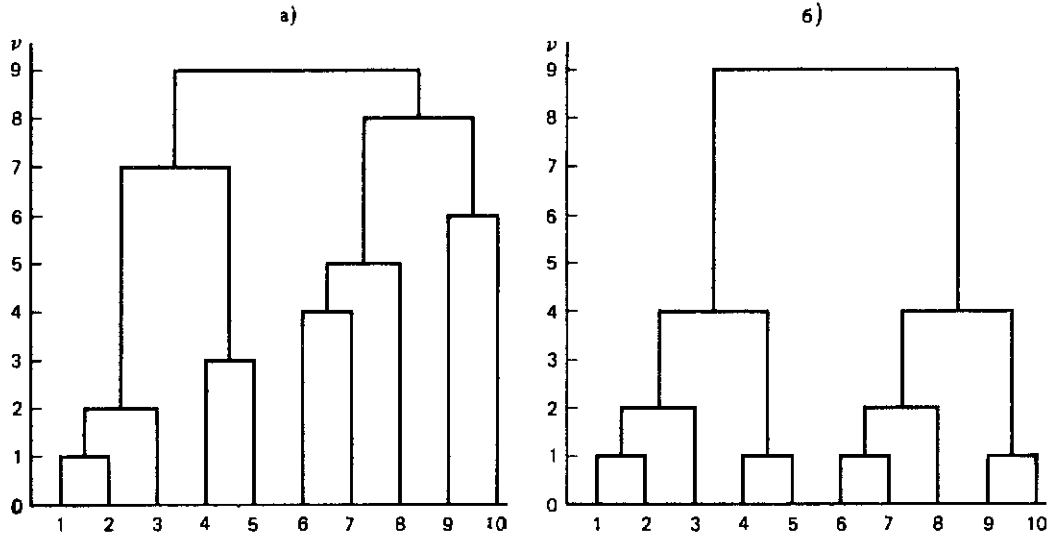


Рис 8 5 Изображение графа иерархии S множества $X=\{x_1, ..., x_{10}\}$ из примера 8.3
а) $\nu(S)=k$, где k — номер шага, на котором S включается в s; *б*) $\nu(S)=|S|-1$

деть, но очень наглядно распределение тех же классов S_i по числу элементов в них, хотя речь идет об одной и той же иерархии s.

Большое значение для визуального анализа иерархии s имеет такое построение ее графа, при котором множество $S \in s$ изображается точкой с ординатой $\nu(S)$, соответствующей значению $\rho(S_1, S_2)$ для множеств S_1 и S_2, из объединения которых оно получено. Такие изображения графа называются *оцифрованными*.

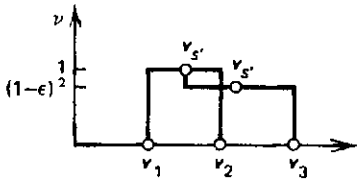


Рис 8 6. Инверсия в изображении графа иерархической классификации множества точек из примера 8 3 Здесь $v_j = \{x_j\}$, $j = 1, 2, 3$, $v_{S'} = \{x_1, x_2\}$, $v_{S''} = \{x_1, x_2, x_3\}$

Возникает задача:

пусть s—бинарная иерархия множества $\mathbf{X}$, полученная агломеративным алгоритмом при помощи меры близости $\rho(S_1, S_2)$. Найти индексацию ν этой иерархии, такую, что

$$\nu(S_1 \cup S_2) = \rho(S_1, S_2). \qquad (8.8)$$

Так как любое множество $S \in s$ однозначно представимо в виде $S' \cup S''$, $S' \cap S'' = \varnothing$ для некоторых S' и S'' из s, то формула (8.8) вместе с условием $\nu(\{X_i\}) = 0$, $X_i \in \mathbf{X}$, $i = 1, \ldots, n$, однозначно задает отображение ν на s. Таким образом, сформулированная задача эквивалентна следующей:

для каких мер близости $\rho(S_1, S_2)$ отображение ν, задаваемое формулой (8.8), является индексацией.

Пример 8.4. Пусть $\mathbf{X} = \{X_1, \ldots, X_n\}$ — набор точек в евклидовом пространстве R^p и $\rho(S_1, S_2) = ||Z_1 - Z_2||^2$, где Z_i — центр класса $S_i \subset \mathbf{X}$, $i = 1, 2$. Тогда отображение ν, задаваемое формулой (8.8), не является индексацией. Действительно, для точек $\{X_1, \ldots, X_{10}\}$ из примера 8.3 имеем в случае $S_1 = \{X_1, X_2\}$ и $S_2 = \{X_3\}$:

$Z_1 = (-1{,}5,\ 1)$, $Z_2 = X_3$ и $\rho(S_1, S_2) = (1 - \varepsilon)^2$, где $\varepsilon > 0$. Таким образом, для $\nu(S) = \rho(S_1, S_2)$ получаем:

$$S_1 \subset S = S_1 \cup S_2, \text{ но } \nu(S) = (1-\varepsilon)^2 < \nu(S_1) = \rho(\{X_1\}, \{X_2\}) = 1$$

— противоречие с определением индексации. Если же мысленно для такого отображения ν изобразим граф иерархии по правилу, описанному выше, то на этом изображении получим следующий фрагмент (рис. 8.6), где вершина, соответствующая объединению двух подмножеств, лежит по оси ор-

динат ниже, чем вершина, соответствующая одному из этих подмножеств. В этом случае говорят, что мера близости имеет инверсии. В § 8.4 приведены условия на меру близости, обеспечивающие отсутствие инверсий у нее. Ряд важнейших мер близости $\rho(S_1, S_2)$ этим условиям удовлетворяет (см. теорему 8.1).

8.4. Приложения общей рекуррентной формулы для мер близости между классами

8.4.1. Расчет матрицы взаимных близостей классов данного уровня иерархии. Рассмотрим более подробно переход от разбиения $S^{(k-1)}$ к разбиению $S^{(k)}$, $k < n-1$, на k-м шаге агломеративного алгоритма.

Имеем $S^{(k-1)} = (S_1^{(k-1)}, \ldots, S_{n-k+1}^{(k-1)})$. Вычислим матрицу взаимных расстояний $\rho^{(k-1)}$ между классами разбиения $S^{(k-1)}$ и найдем пару классов (S', S''), такую, что

$$(S', S'') = \arg \min_{1 \leqslant l < t \leqslant n-k+1} \rho\left(S_l^{(k-1)}, S_t^{(k-1)}\right).$$

Пусть $S' = S_i^{(k-1)}$ и $S'' = S_j^{(k-1)}$, где $i < j$. Тогда положим

$$S_l^{(k)} = S_l^{(k-1)},\ l \neq i,\ j,\ n-k+1,\ S_i^{(k)} =$$
$$= S_i^{(k-1)} \cup S_j^{(k-1)},\ S_j^{(k)} = S_{n-k+1}^{(k-1)}.$$

Следовательно, для получения матрицы взаимных расстояний $\rho_l^{(k)}$ можно воспользоваться матрицей $\rho^{(k-1)}$, вычислив дополнительно только расстояния $\rho(S_i^{(k)}, S_l^{(k)})$, $l \neq i$ при фиксированном i. Так как $S_i^{(k)} = S_i^{(k-1)} \cup S_j^{(k-1)}$, то для этого оказывается полезной рекуррентная формула Жамбю [248]:

$$\rho(S, S' \cup S'') = a_1 \rho(S, S') + a_2 \rho(S, S'') + a_3 \rho(S', S'') +$$
$$+ a_4 \nu(S) + a_5 \nu(S') + a_6 \nu(S'') + a_7 |\rho(S, S') - \rho(S, S'')|, \quad (8.9)$$

обобщающая формулу Ланса и Вильямса (5.9).

8.4.2. Условия на меру близости, обеспечивающие отсутствие инверсий. Эти условия дает результат, полученный Г. Миллиганом [248].

Т е о р е м а 8.1. Пусть s — иерархия на множестве $\mathbf{X}$, полученная при помощи меры близости $\rho(S_1, S_2)$, для которой верна формула (8.9). Тогда, если $a_1 + a_2 + a_3 \geqslant 1$, $a_j \geqslant 0$ для $j = 1, 2, 4, 5, 6$ и $a_7 \geqslant -\min(a_1, a_2)$, то отображение ν: $s \to R^1$, задаваемое формулой $\nu(S_1 \cup S_2) =$

$= \rho(S_1, S_2)$ и условием $\nu(\{X_i\}) = 0$, $i = 1, \ldots, n$, является индексацией.

Пример 8.5. Пусть $\mathbf{X} = \{X_1, \ldots, X_n\}$ — набор точек в евклидовом пространстве R^p и

$$\rho(S_1, S_2) = \sum_{X \in S_1 \cup S_2} \|X - Z\|^2 - \sum_{X \in S_1} \|X - Z_1\|^2 - \sum_{X \in S_2} \|X - Z_2\|^2, \quad (8.10)$$

где Z, Z_1 и Z_2 — центры классов $S_1 \cup S_2$, S_1 и S_2. Тогда верна формула

$$\rho(S, S' \cup S'') = \frac{n+n'}{n+n'+n''}\rho(S, S') + \frac{n+n''}{n+n'+n''} \times$$

$$\times \rho(S, S'') - \frac{n}{n+n'+n''}\rho(S', S'')$$

(ср. с формулой (5.10), где $n = |S|$, $n' = |S'|$ и $n'' = |S''|$. В обозначениях формулы (8.9) имеем:

$$a_1 = \frac{n+n'}{n+n'+n''}, \quad a_2 = \frac{n+n''}{n+n'+n''}, \quad a_3 = -\frac{n}{n+n'+n''}, \quad a_k = 0, \quad k = 4, \ldots, 7.$$

Имеем $a_1 + a_2 + a_3 = 1$, т. е. все условия теоремы Г. Милигана выполняются и поэтому мера близости (8.10) не имеет инверсий.

Замечание. Из (8.10) следует, что

$$\rho(S_1, S_2) = \frac{n_1 n_2}{n_1 + n_2}\|Z_1 - Z_2\|^2,$$

где $n_1 = |S_1|$, $n_2 = |S_2|$. Таким образом, введение множителя $\frac{n_1 n_2}{n_1 + n_2}$ позволяет исправить инверсию меры близости $\rho(S_1, S_2) = \|Z_1 - Z_2\|^2$ из примера 8.4, которая удовлетворяет формуле (8.9) с параметрами:

$$a_1 = \frac{n_1}{n_1 + n_2}, \quad a_2 = \frac{n_2}{n_1 + n_2}, \quad a_3 = -\frac{n_1 n_2}{(n_1 + n_2)^2}, \quad a_k = 0, \quad k = 4, \ldots, 7.$$

Здесь $a_1 + a_2 + a_3 = 1 - \frac{n_1 n_2}{(n_1 + n_2)^2} < 1$.

8.4.3. Алгоритм гибкой стратегии иерархической классификации. Формула Жамбю (8.9) позволяет обобщить *гибкую стратегию* Ланса и Вильямса [150].

Пусть исходная информация о классифицируемых объектах представлена в форме матрицы взаимных расстояний

$\rho = (\rho_{ij})$, $1 \leqslant i, j \leqslant n$. В алгоритмах гибкой стратегии расстояние между классами $\rho(S_1, S_2)$, где $|S_1| \cdot |S_2| > 1$, заранее не фиксируется, как это делается обычно, а настраивается в ходе классификации. Положим $\rho^{(0)} = \rho$. Пусть, по индуктивному предположению, известны матрица взаимных расстояний $\rho^{(k-1)}$ между классами разбиения $S^{(k-1)}$, $k = 1, \ldots, n-2$, и индексирующее отображение ν: $S^{(k-1)} \to R^1$. Напомним, что ν на $S^{(0)}$ — нулевое отображение. Тогда, подобрав любым способом параметры $a_1, \ldots, a_7$, удовлетворяющие теореме 8.1, получаем возможность при помощи формулы Жамбю вычислить матрицу взаимных расстояний $\rho^{(k)}$ между классами разбиения $S^{(k)}$, а при помощи формулы $\nu(S_1 \cup S_2) = \rho(S_1, S_2)$ — индексирующее отображение ν: $S^{(k)} \to R^1$.

8.4.4. Процедуры иерархической классификации, использующие пороговые значения. Как уже отмечалось выше, трудности реализации классических алгоритмов иерархической классификации (см. п. 8.2.2) быстро растут с ростом числа n классифицируемых объектов. Это объясняется тем, что на k-м шаге алгоритма для каждого $k = 1, \ldots, n-1$ приходится искать минимальный элемент в матрице взаимных расстояний $\rho^{(k-1)}$, т. е. находить минимальный элемент в массиве из $N_{k-1} = (n-k+1)(n-k)/2$ чисел. Ясно, что минимальный элемент в матрице $\rho^{(k-1)}$ достаточно искать среди ее элементов, не превосходящих некоторый порог c, а массив таких элементов при соответствующем выборе порога c содержит элементов намного меньше, чем N_{k-1}. В связи с этим разрабатываются процедуры иерархической классификации, использующие пороговые значения. Общая схема подобных процедур следующая: задается или формируется в процессе классификации последовательность порогов $c_1 < c_2 < \ldots < c_t$. На первом этапе алгоритма попарно объединяются элементы, а затем и классы, меры близости которых не превосходят c_1. На втором этапе — c_2 и т. д., пока все элементы не объединятся в один класс. Эффективность таких процедур существенно зависит от внутренней структуры исследуемого множества объектов, от выбора последовательности пороговых значений $c_1, \ldots, c_t$ и меры близости $\rho(S_1, S_2)$ между классами. Сравнительно недавно было показано, что если мера близости $\rho(S_1, S_2)$ обладает свойством редуктивности (см. определение 8.7), то процедуры иерархической классификации, использующие пороговые значения, позволяют построить точно такую иерархию, что и классические агломеративные процедуры, но работают они существенно быстрее последних [249]. Этим вопросам посвящен § 8.5.

Опишем свойство редуктивности мер близости и способ его проверки.

Пусть $s = \{S^{(0)} \subset \ldots \subset S^{(k-1)} \subset S^{(k)} \subset \ldots\}$ — бинарная иерархия, полученная агломеративным алгоритмом при помощи меры близости $\rho(S_1, S_2)$ и $S_{i_*}^{(k-1)}$, $S_{j_*}^{(k-1)}$ — пара классов, которая объединяется в один класс на k-м уровне иерархии.

Определение 8.7. Мера близости $\rho(S_1, S_2)$ обладает свойством *редуктивности* (является *редуктивной*), если для любого класса $S_l \in S^{(k-1)}$, $k = 1, \ldots, n-1$, и порога $c > 0$, такого, что $\rho(S_{i_*}^{(k-1)}, S_{j_*}^{(k-1)}) < c$, из условий $\rho(S_l, S_{i_*}^{(k-1)}) \geqslant c$ и $\rho(S_l, S_{j_*}^{(k-1)}) \geqslant c$ вытекает, что $\rho(S_l, S_{i_*}^{(k-1)} \cup S_{j_*}^{(k-1)}) \geqslant c$.

Для мер близости $\rho(S_1, S_2)$, удовлетворяющих формуле Жамбю (8.9), свойство редуктивности проверяется при помощи следующего теоретического результата:

Теорема 8.2 (Диде). Если параметры $a_1, \ldots, a_7$ меры близости $\rho(S_1, S_2)$ удовлетворяют условиям $a_1 + a_2 + \left(\frac{a_3 - |a_3|}{2}\right) \geqslant 1$, $a_7 \geqslant -\min(a_1, a_2)$, $a_j \geqslant 0$, $j = 1, 2, 4, 5, 6$, то мера близости $\rho(S_1, S_2)$ является редуктивной.

Сравнивая утверждения теорем 8.1 и 8.2, получаем, что для каждой редуктивной меры близости $\rho(S_1, S_2)$ отображение $\nu : s \to R^1 : \nu(S_1 \cup S_2) = \rho(S_1, S_2)$ является индексацией соответствующей ей иерархии s.

Пример 8.6. Для меры близости $\rho(S_1, S_2)$ из примера 8.5 имеем: $a_1 + a_2 + \frac{a_3 - |a_3|}{2} = 1$, т. е. приращение внутриклассового разброса при объединении классов является редуктивной мерой близости между классами. Другие важные примеры рассмотрены в § 8.5.

8.5. Быстрый алгоритм иерархической классификации

Опишем сначала основной блок алгоритма.

Пусть $\mathbf{X} = \{X_1, \ldots, X_n\}$ и $\rho(S_1, S_2)$ — некоторая мера близости между подмножествами из $\mathbf{X}$. Выберем порог c.

Из матрицы взаимных расстояний $\rho_{\mathbf{X}} = (\rho_{ij} = \rho(X_i, X_j))$ выберем массив $W_c = (\rho_{ij} : \rho_{ij} < c)$. Допустим, что $W_c \neq \varnothing$.

Шаг А.

Найдем $\rho_{i_* j_*} = \min\limits_{W_c} \rho_{ij}$.

Объединим точки X_{i_*} и X_{j_*} в один класс $\{X_{i_*}, X_{j_*}\}$, которому присвоим обозначение X_{n+1}. Положим

$$\mathbf{X}^1 = \{X_1, \ldots, X_{i_*}, \ldots, X_{j_*}, \ldots, X_n, X_{n+1}\},$$

где $\cdot$ — символ удаления элемеитов.

Сформируем массив W_c^1 из матрицы взаимных расстояний множества $\rho_{\mathbf{X}1}$, включив в него:

ρ_{ij}, если $\rho_{ij} \in W_c$ и $\{i, j\} \cap \{i_*, j_*\} = \varnothing$;

$\rho_{i,n+1} = \rho(X_i, X_{n+1})$, если $\rho_{ii_*} \in W_c$ или $\rho_{ij_*} \in W_c$,

причем $\rho_{i,\,n+1} < c$.

Таким образом получаем пару $(\mathbf{X}^1, W_c^1)$. Если $W_c^1 \neq \varnothing$, то описаиный шаг А можно повторить.

Итогом работы рассматриваемого блока алгоритма является последовательность пар

$$(\mathbf{X}, W_c), \ (\mathbf{X}^1, W_c^1), \ \ldots, \ (\mathbf{X}^l, W_c^l), \ 1 \leqslant l \leqslant n-1,$$

где $\mathbf{X}^k$ — миожество, полученное из $\mathbf{X}^{k-1}$, $k \geqslant 1$, объединеиием двух каких-либо элементов, W_c^k — массив, сформированный из матрицы взаимных расстояиий множества $\mathbf{X}^k$ и W_c^l — пустой массив.

Схема алгоритма

1. Задаем множество $\mathbf{X} = \{X_1, \ldots, X_n\}$ и меру близости $\rho(S_1, S_2)$.
2. Выбираем значение порога c_0 и формируем массив W_c. Например, c_0 выбирается так, чтобы $q_1 |\mathbf{X}| \leqslant |W_c| \leqslant q_2 |\mathbf{X}|$, где q_1, q_2 — числовые параметры.
3. Применяем к $(\mathbf{X}, W_c)$ процедуру осиовного блока алгоритма. В результате находим множество $\mathbf{X}^l$, для которого $W_c^l = \varnothing$. По построению элементам множества $\mathbf{X}^l$ соответствуют непересекающиеся классы из $\mathbf{X}$, т. е. $\mathbf{X}^l$ — разбиение множества $\mathbf{X}$.
4. Если $|\mathbf{X}^l| > 1$, то возвращаемся к шагу 1, заменив $\mathbf{X}$ на $\mathbf{X}^l$. (В алгоритмах гибкой стратегии заменяется и мера близости.) Если $|\mathbf{X}^l| = 1$, то заканчиваем работу алгоритма.

Итогом работы алгоритма является бинарная иерархия на $\mathbf{X}$, но в общем случае эта иерархия ие совпадает с иерархией, получаемой классической агломеративной процедурой при помощи меры близости $\rho(S_1, S_2)$. Соответствующий при-

мер нетрудно привести для меры близости $\rho(S_1, S_2) = ||Z_1 - Z_2||^2$, где Z_1, Z_2 — центры классов.

В то же время, как уже отмечалось, для редуктивных мер близости (см. определение 8.7) описанная процедура представляет собой ускоренный вариант классической агломеративной процедуры.

Приведем важнейшие редуктивные меры близости. Проверка опирается на теорему 8.2 и проводится в терминах параметров $a_1, \ldots, a_7$ общей рекуррентной формулы для мер близости.

Положим $a_1 + a_2 + \left(\frac{a_3 - |a_3|}{2}\right) = b$ и $\min(a_1, a_2) = -d$. Тогда, по теореме 8.2, если параметры $a_1, \ldots, a_7$ таковы, что $b \geqslant 1$, $a_j \geqslant 0$, $j = 1, 2, 4, 5, 6$ и $d \leqslant a_7$, то определяемая ими мера близости $\rho(S_1, S_2)$ редуктивна.

Для краткости указываем только наименование меры близости:

1) «ближайший сосед» (5.4):

$$a_1 = a_2 = -a_7 = \frac{1}{2}, \quad a_j = 0, \quad 3 \leqslant j \leqslant 6,$$

$$b = 1, \quad d = -\frac{1}{2} = a_7;$$

2) «дальний сосед» (5.5):

$$a_1 = a_2 = a_7 = \frac{1}{2}, \quad a_j = 0, \quad 3 \leqslant j \leqslant 6,$$

$$b = 1, \quad d = -\frac{1}{2} < a_7;$$

3) средней связи (5.7):

$$a_1 = \frac{n_1}{n_1 + n_2}, \quad a_2 = \frac{n_2}{n_1 + n_2}, \quad a_j = 0, \quad 3 \leqslant j \leqslant 7,$$

$$b = 1, \quad d = -\frac{\min(n_1, n_2)}{n_1 + n_2} < 0;$$

4) приращение статистического разброса при объединении классов (см. примеры (8.5 и 8.6)).

ВЫВОДЫ

1. Введены понятия иерархии s на множестве объектов $\mathbf{X} = \{X_1, \ldots, X_n\}$ и графа иерархии.
2. Общая постановка задачи иерархической классификации состоит в требовании построить иерархию s на исследуе-

мой совокупности объектов, отражающую наличие однородных, в определенном смысле, классов и взаимосвязи между этими классами.

3. В основе алгоритмов иерархической классификации лежит тот или иной критерий качества $Q(S_1, \ldots, S_k)$ разбиения множества S на подмножества $S_1, \ldots, S_k$. Обычно используются бинарные алгоритмы ($k=2$). В этом случае $Q(S_1, S_2)$ имеет смысл близости $\rho(S_1, S_2)$ между множествами S_1 и S_2. Мера близости $\rho(S_1, S_2)$ либо фиксирована, либо меняется (настраивается) в ходе алгоритма. В последнем случае говорят о гибкой (адаптивной) стратегии классификации.

4. Алгоритмы иерархической классификации бывают:

а) дивизимные, результатом которых является система вложенных разбиений $S^{(0)} \supset S^{(1)} \supset \ldots \supset S^{(n-1)}$ множества $\mathbf{X}$, где $S^{(0)} = \mathbf{X}$ и $S^{(k)} = (S_1^{(k)}, \ldots, S_{k+1}^{(k)})$ — разбиение $\mathbf{X}$ на непересекающиеся классы, называемое разбиением k-го уровня. Переход с k-го уровня на $(k+1)$ осуществляется разбиением некоторого класса $S_l^{(k)}$ на подклассы S_1^*, S_2^*, где

$$(S_1^*, S_2^*) = \arg \max_{S_1 \cup S_2 = S_l^{(k)}} \rho(S_1, S_2);$$

б) агломеративные, результатом которых является система вложенных разбиений $S^{(0)} \subset S^{(1)} \subset \ldots \subset S^{(n-1)}$ множества $\mathbf{X}$, где $S^{(0)} = (X_1, \ldots, X_n)$ и $S^{(k)} = (S_1^{(k)}, \ldots, S_{n-k}^{(k)})$ — разбиение k-го уровня. Переход с k-го уровня на $(k+1)$-й осуществляется объединением пары классов (S_1^*, S_2^*), где

$$(S_1^*, S_2^*) = \arg \min_{\substack{S_1 \neq S_2 \\ S_1, S_2 \in S^{(k)}}} \rho(S_1, S_2).$$

5. Существенным преимуществом алгоритмов иерархической классификации является возможность изображения на плоскости графа полученной иерархии, называемого графом классификации. Имеется большая свобода в построении на плоскости данного графа классификации. Изучая различные изображения, согласованные со структурой этого графа, можно получить ряд тонких результатов об исследуемом множестве объектов.

6. Описана общая схема построения на плоскости графа классификации. В основе построения лежит понятие индексации иерархии, которая представляет собой отображение, ставящее в соответствие каждому классу S число $\nu(S)$ таким образом, что $\nu(S) = 0$ тогда и только тогда,

когда S состоит из одного элемента, и $\nu(S') \leqslant \nu(S)$, как только $S' \subset S$.

7. Для визуального анализа иерархии s большое значение имеет построение на плоскости такого изображения ее графа, при котором классу $S \in s$ соответствует точка с ординатой $\nu(S) = \rho(S_1, S_2)$, где S_1 и S_2 — классы, из объединения которых получен класс S. Такие изображения иазываются оцифрованными.
8. Важнейшие меры близости $\rho(S_1, S_2)$ описываются общей рекуррентной формулой (8.9).

В терминах параметров $a_1, \ldots, a_7$ из (8.9)

а) даны условия, достаточные для существования оцифрованного изображения графа классификации, соответствующего мере близости с этими параметрами;

б) описана общая схема построения алгоритмов гибкой стратегии иерархической классификации;

в) описана общая схема процедур иерархической классификации, использующих пороговые значения, но дающих тот же результат, что и классические агломеративные процедуры. В рамках этой схемы описан быстрый алгоритм иерархической классификации, позволяющий получать иерархическую классификацию больших массивов данных на ЭВМ средней мощности, в частности на персональных компьютерах.

Глава 9. ПРОЦЕДУРЫ КЛАСТЕР-АНАЛИЗА И РАЗДЕЛЕНИЯ СМЕСЕЙ ПРИ НАЛИЧИИ АПРИОРНЫХ ОГРАНИЧЕНИЙ

9.1. Разделение смесей при наличии неполных обучающих выборок

Здесь и далее в главе рассматриваются процедуры кластер-анализа и разделения смесей распределений, когда у исследователя имеется некоторая априорная информация относительно желаемой классификации, задаваемая в виде тех или иных ограничений.

Иногда возникает ситуация, когда исследователю известна принадлежность некоторых объектов из матрицы данных $\mathbf{X}$ к некоторым компонентам смеси или кластерам (классам). Дальше будем считать без ограничения общности, что имеются обучающие выборки (ОВ) $\{\mathbf{X}\}_j^{\nu_j} = (\overline{1, l})$ для l первых классов и объем такой выборки ν_j. Суммарный объем таких выбо-

рок $\nu = \Sigma \nu_j \ll n$ и ие позволяет воспользоваться процедурами дискриминантиого анализа. Количество ОВ (l) может быть меньше количества выделяемых классов k.

9.1.1. Модификация ЕМ-алгоритма. ЕМ-алгоритм для оценки параметров смеси распределений описан в § 6.4. Этот алгоритм носит итерационный характер, на каждом шаге t, в частности, пересчитываются вероятности принадлежности i-го объекта X_i к j-му классу по формуле (6.9)

$$g_{ij}^{(t)} = p_j f\left(X_i;\ \Theta_j^{(t-1)}\right) \Big/ \sum_{j=1}^{k} p_j f\left(X_i;\ \Theta_j^{(t-1)}\right). \qquad (9.1)$$

Модификация алгоритма при наличии неполных ОВ состоит в том, что для объектов, которые в них содержатся, значения $g_{ij}^{(t)}$ корректируются следующим образом [66]: если объект X_i принадлежит ОВ для r-го класса, то $g_{ij}^{(t)} = 0$ ($j \neq r$, $j = \overline{1, k}$) и $g_{ir}^{(t)} = 1$.

Эффективность использования неполных ОВ весьма велика. Имеются примеры, когда использование ОВ, составляющих примерно 10 % исходной выборки, приводило к резкому улучшению результата разделения смеси [1].

9.1.2. Разделение смеси с неизвестным числом классов. Рассмотрим случай смеси нормальных распределений с равными матрицами ковариаций, число компонентов k которой неизвестно. Кроме того, имеются неполные ОВ, так же как и в п. 9.1.1.

Вычислительная процедура состоит из следующих шагов [66].

Шаг 1. Вычисляются оценки векторов средних значений $\overline{X_j}$ ($j = \overline{1, l}$) и общей матрицы ковариаций $\Sigma_{\nu-l}$ по неполным ОВ. Нижний индекс указывает число степеней свободы, соответствующее оценке матрицы ковариаций. Далее для измерения расстояния между объектами X_i и X_r используется расстояние Махаланобиса

$$d^2(X_i,\ X_r) = (X_i - X_r)'\ \widehat{\Sigma}_{\nu-l}^{-1}\,(X_i - X_r). \qquad (9.2)$$

Пусть теперь $\mathbf{h}$ — вектор размерности n, у которого i-я компонента равна номеру класса для объекта X_i. Приравниваем к нулю компоненты $\mathbf{h}$, а объектам из ОВ присваиваем соответствующие номера. Текущее значение числа клас-

---

[1] *Hosmer D. W. Jr.* A comparasion of Iterative Maximum Likehood Estimates of the Parameters of a Mixture of Two Normal Distribution under Three Different Types of Sample //Biometrics. — 1973. — Vol. 29.—P. 761—760.

сов m полагается равным l. Значение счетчика числа классифицированных объектов $\nu = \sum_{j=1}^{l} \nu_j$, где ν_j — объемы ОВ.

Шаг 2. Обнуляются счетчики числа классификаций объектов η и числа случаев образования новых классов ξ.

Проведем последовательный просмотр неклассифицированных объектов, т. е. объектов X_i, для которых $h_i = 0$.

Пусть X_i — такой объект. Тогда вычисляются расстояния от X_i до центров уже образованных классов $d^2(X_i, \overline{X}_j)$, величины $t_j^2 = \frac{\nu_j(\nu - m - p + 1)}{p(\nu_j - 1)(\nu - m)} d^2(X_i, \overline{X}_j)$ и значения $F(t_j^2)$ функции F-распределения с p и $\nu - m - p + 1$ степенями свободы от t_j^2. Вычисляется $f_j = 1 - F(t_j^2)$. При сделанных допущениях (нормальность, равные матрицы ковариаций) величина f_j равна вероятности реализации расстояния от X_i до $\overline{X}_j$, большего или равного t_j^2 при условии, что X_i действительно принадлежит j-му классу. Пусть теперь

$$f_{\max} = \max_{1 \leqslant j \leqslant m} f_j, \quad j_{\max} = \arg\max_{1 \leqslant j \leqslant m} f_j, \text{ т. е. } f_{\max} = f_{j_{\max}}.$$

Относительно объекта X_i принимается одно из трех решений:

1) если $f_{\max} > c_1$, то объект X_i относится к классу с номером $j_{\max}$ и проводится корреляция оценки $\overline{X}_{j\max}$ и $\widehat{\Sigma}$:

$$\overline{X}_{j_{\max}}^{\text{нов}} = (\nu_j X_{j_{\max}} + X_i)/(\nu_{j_{\max}} + 1); \tag{9.3}$$

$$\widehat{\Sigma}^{\text{нов}} = \widehat{\Sigma} + \frac{\nu + 1}{\nu^2}(X_i - \overline{X}_{j_{\max}}^{\text{нов}})(X_i - \overline{X}_{j_{\max}}^{\text{нов}})'. \tag{9.4}$$

Используя формулу Бартлетта [129], получаем скорректированную обратную матрицу

$$\left(\widehat{\Sigma}^{\text{нов}}\right)^{-1} = \widehat{\Sigma}^{-1} - \frac{\frac{\nu+1}{\nu^2}\widehat{\Sigma}^{-1}(X_i - \overline{X}_{j_{\max}}^{\text{нов}})(\widehat{\Sigma}^{-1}(X_i - \overline{X}_{j_{\max}}^{\text{нов}}))'}{1 + \frac{\nu+1}{\nu^2}(X_i - \overline{X}_{j_{\max}}^{\text{нов}})'\widehat{\Sigma}^{-1}(X_i - \overline{X}_{j_{\max}}^{\text{нов}})}; \tag{9.5}$$

$$\nu_{j_{\max}} = \nu_{j_{\max}} + 1; \quad \nu = \nu + 1; \quad h_i = j_{\max};$$

$$\eta = \eta + 1;$$

2) если $f_{\max} < c_2$, то считается, что объект X_i принадлежит некоторому новому $(m+1)$-му классу; счетчик числа классов m увеличивается на 1 ($m = m + 1$) и полагается, $\overline{X}_m = X_i$, $h_i = m$; $\xi = \xi + 1$; $\nu = \nu + 1$;

3) если выполняются неравенства $c_2 \leqslant f_{\max} \leqslant c_1$, то никаких действий не проводится.

Если просмотр объектов не окончен, то переходим к просмотру следующего объекта.

Шаг 3. Проверяется, все ли объекты расклассифицированы, т. е. равенство $\nu = n$. Если оно выполняется, то производится переход на шаг 5, в противном случае на шаг 4.

Шаг 4. Проверяются значения счетчиков η и ξ. Если хотя бы один из счетчиков не равен нулю, то переходят на шаг 2. Если одновременно $\eta = 0$ и $\xi = 0$, то, следовательно, на шаге 2 не было образовано ни одного нового класса и не было классификации объектов. Поэтому проводится уменьшение порога c_1 на величину δc_1 и увеличение порога c_2 на величину δc_2, т. е. $c_1 = c_1 - \delta c_1$, $c_2 = c_2 + \delta c_2$. Таким образом, увеличиваются возможности классификации объектов и образования новых классов (принятые при реализации алгоритма значения $\delta c_1 = 0{,}01$ и $\delta c_2 = 0{,}001$). Производится переход на шаг 2.

Шаг 5. Проводится реклассификация исходной совокупности объектов X так же, как на шаге 2, но при $c_1 = 0$ и без пересчета оценок статистических характеристик $\overline{X}_j$, $\widehat{\Sigma}$. Полученная классификация считается окончательной.

9.2. Классификация при ограничениях на связи между объектами

Довольно типичной является ситуация, когда исследователь желает получить разбиение совокупности объектов $O_1, \ldots, O_n$ на классы, согласованное с матрицей ограничений $\mathbf{B} = (b_{ij})$, $1 \leqslant i, j \leqslant n$, где $b_{ij} = b_{ji}$, $b_{ii} = 1$ и $b_{ij} = 0$, $i \neq j$, если объекты O_i и O_j, по априорным сведениям, являются разнородными, и $b_{ij} = 1$, если таких сведений об объектах O_i и O_j нет. Например, в задачах формирования нескольких экспертных комиссий для анализа некоторой производственно-экономической системы объектами группировки являются эксперты, описываемые их профессиональными показателями, но при формировании комиссий желательно учитывать особенности взаимоотношений между экспертами. В ряде задач, где объекты описываются

даниями измерений, проб и т. п., при классификации часто необходимо учитывать качественную однородность этих объектов.

Приведем примеры построения матриц ограничений $\mathbf{B} = (b_{ij})$, $1 \leqslant i, j \leqslant n$.

Пример 9.1. Пусть $O^t \subset O = (O_1, \ldots, O_n)$ — подмножество объектов, объединенных общностью какого-либо показателя π_t, скажем, в задачах социологии, демографии и т. п. Допустим, что набор показателей π_t, $1 \leqslant t \leqslant T$ позволяет представить O в виде объединения $\bigcup_{t=1}^{T} O^t$, вообще говоря, пересекающихся подмножеств O^t. Тогда для пары объектов O_i и O_j положим $b_{ij} = 0$, если не существует ни одного показателя π_t, объединяющего их, и $b_{ij} = 1$, если $\{O_i, O_j\} \subset O^t$ для некоторого t.

Пример 9.2. Пусть исследуемый объект O_i описывается точкой $X_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \in R^p$, причем набор признаков $x_i^1, \ldots, x^{(p)}$, таков, что для некоторого порогового значения c можно, априори, сделать вывод: если $||X_i - X_j|| > c$, то объекты O_i и O_j не являются однородными. Тогда положим $b_{ij} = 0$, если $||X_i - X_j|| > c$, и $b_{ij} = 1$, если $||X_i - X_j|| \leqslant c$.

Пример 9.3. Пусть объект O_i описывается парой (X_i, Z_i), где $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})$ — вектор признаков, поддающихся измерению, а $Z_i = (z_i^{(1)}, \ldots, z_i^{(q)})$ — вектор признаков, для которых не существует объективно обусловленной шкалы, скажем, $z_i^{(q)}$ — выраженное в баллах мнение q-го эксперта об i-м объекте. Допустим, что существует мера близости $\rho'(Z_i, Z_j)$, такая, что для некоторого порогового значения c из условия $\rho'(Z_i, Z_j) > c$ вытекает, что объекты O_i и O_j заведомо не являются однородными. Тогда положим $b_{ij} = 0$, если $\rho'(Z_i, Z_j) > c$, и $b_{ij} = 1$, если $\rho'(Z_i, Z_j) \leqslant c$.

Итак, пусть имеется совокупность объектов $O = (O_1, \ldots, O_n)$, исходная информация о которых представлена либо в форме матрицы объект — свойство $\mathbf{X} = (x_i^{(j)})$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, и матрицы ограничений $\mathbf{B} = (b_{ij})$, $1 \leqslant i, j \leqslant n$, где $b_{ij} = 0$ или 1, либо в форме матрицы $\rho = (\rho_{ij})$, $1 \leqslant i, j \leqslant n$ взаимных близостей между объектами, причем в этой матрице пропущены элементы ρ_{ij}, где (i, j) — пары индексов, для которых $b_{ij} = 0$.

Опишем схему соответствующего агломеративного алгоритма иерархической классификации.

Схема алгоритма

1. Выберем меру близости $\rho(S_1, S_2)$ между подмножествами исследуемой совокупности. Подадим на вход алгоритма разбиение $S^0 = (S_1^0, \ldots, S_n^0)$ на одноточечные классы $S_i^0 = \{X_i\}$ и матрицу ограничений $\mathbf{B} = (b_{ij})$.

2. Допустим, что на m-м шаге имеется разбиение $S^{m-1} = (S_1^{m-1}, \ldots, S_{n-m+1}^{m-1})$, $1 \leqslant m \leqslant n-1$, и матрица ограничений $\mathbf{B}^{m-1} = (b_{ij}^{m-1})$, $1 \leqslant i, j \leqslant n-m+1$.
Найдем

$$(i_*, j_*) = \arg\min_{(i,j)} \{\rho(S_i^{m-1}, S_j^{m-1}) : i < j,\ b_{ij}^{m-1} = 1\}.$$

Объединим классы $S_{i_*}^{m-1}$ и $S_{j_*}^{m-1}$ в один класс и получим разбиение $S^m = (S_1^m, \ldots, S_{n-m}^m)$, где

$$S_t^m = S_t^{m-1}, \text{ если } t \neq i_*, j_*, n-m+1;$$

$$S_{i_*}^m = S_{i_*}^{m-1} \cup S_{j_*}^{m-1} \text{ и } S_{j_*}^m = S_{n-m+1}^{m-1}.$$

Далее, получим матрицу ограничений $\mathbf{B}^m = (b_{ij}^m)$, где

$b_{ij}^m = b_{ij}^{m-1}$, если $\{i, j\} \cap \{i_*, j_*\} = \varnothing$;

$$b_{i_* j}^m = \begin{cases} 0, \text{ если } b_{i_* j}^{m-1} + b_{j_* j}^{m-1} = 0; \\ 1, \text{ если } b_{i_* j}^{m-1} + b_{j_* j}^{m-1} \neq 0, \ j \neq j_*; \end{cases}$$

$$b_{i_* j_*}^m = \begin{cases} 0, \text{ если } b_{i_*, n-m+1}^{m-1} + b_{j_*, n-m+1}^{m-1} = 0; \\ 1, \text{ если } b_{j_*, n-m+1}^{m-1} + b_{j_*, n-m+1}^{m-1} \neq 0; \end{cases}$$

$b_{i j_*}^m = b_{i, n-m+1}^{m-1}$, если $i \neq i_*$.

3. Если $\mathbf{B}^m$ — единичная матрица, то объявляем разбиение S^m итоговой классификацией. Если вне диагонали матрицы $\mathbf{B}^m$ имеются ненулевые элементы, то возвращаемся к шагу 2, заменив m на $m+1$.

Результатом работы алгоритма является последовательность разбиений $S^0 \subset \ldots \subset S^k$, где $0 \leqslant k \leqslant n-1$, причем каждое разбиение $S^q = (S_1^q, \ldots, S_{n-q}^q)$ согласовано с матрицей ограничений $\mathbf{B}$ в следующем смысле: если объекты O_i и O_j попадают в один класс, скажем S_t^q, $1 \leqslant t \leqslant n-1$, то в этом классе обязательно содержится цепь объектов O_i, $O_{i_1}, \ldots, O_{i_{m-1}}, O_j$, $m \geqslant 1$, такая, что $b_{i, i_1} \ldots b_{i_{m-1}, j} = 1$.

Все результаты об агломеративных алгоритмах иерархической классификации (см. гл. 8) естественным образом распространяются на описанный выше алгоритм классификации при ограничениях на связи между объектами.

9.3. Классификация на графах

Опишем методы и алгоритмы классификации, основанные на представлении исходной информации о классифицируемых объектах в виде графа близости $G = (V, E)$, вершины $v_1, \ldots, v_n \in V$ которого соответствуют объектам $O_1, \ldots, O_n$, а ребра $e_{ij} \in E$, $i \neq j$, соединяющие вершины v_i и v_j — неупорядоченным парам (O_i, O_j), $i \neq j$, т. е. $e_{ij} = e_{ji}$. Длина ребра e_{ij} считается равной ρ_{ij} для выбранной меры близости между объектами O_i и O_j.

В изложении будем существенно опираться на работу Д. В. Матулы из [83, с. 83—111].

9.3.1. Основные понятия и определения. Предварительные сведения из теории графов приведены в [12, п. 4.2.1].

Определение 9.1. Граф $G = (V, E)$, где $V = \{v_1, \ldots, v_n\}$, $E = (e_{ij})$, называется *полным*, если любые его две вершины v_i и v_j соединены ребром $e_{ij} \in E$.

Например, граф близости $G = G(O)$ совокупности объектов O является полным. В задачах классификации при наличии ограничений на связи между объектами (см. § 9.2) используются неполные графы (**B**-графы близости, где $\mathbf{B} = (b_{ij})$), полученные из полного графа близости удалением ребер e_{ij} при условии, что $b_{ij} = 0$.

Определение 9.2. Пусть $G = (V, E)$ — некоторый граф. Вектором инцидентности вершины v_i называется вектор $w_i = (w_{i1}, \ldots, w_{in})$, где $w_{ij} = 1$, если $e_{ij} \in E$, и $w_{ij} = 0$, если $e_{ij} \notin E$[1].

Степенью вершины v_i в графе G называется число $d_i(G) = d_i = \sum_{j=1}^{n} w_{ij}$. Ясно, что граф G полный тогда и только тогда, когда степень любой его вершины равна $n - 1$.

Определение 9.3. Подграф $G' \subseteq G$ называется максимальным по отношению к некоторому свойству F (F-максимальным), если G' обладает свойством F и в G не существует подграфа G'', обладающего свойством F, такого, что $G' \subset G'' \subset G$, $G' \neq G'' \neq G$.

Пример 9.4. Пусть $O' \subset O = \{O_1, \ldots, O_n\}$ и $V' \subset V$ — соответствующее подмножество вершин графа близости $G(O) = (V, E)$. Тогда среди всех подграфов с множеством вершин V' максимальным является граф близости $G(O')$.

---

[1] Обратим внимание, что $w_{ii} = 0$ для всех i, т. е., как и в [12], рассматриваем только простые графы, у которых ребра, соединяющие вершину с собой, отсутствуют.

Напомним, что граф $G = (V, E)$ называется связанным, если любые две его вершины v_i, v_j можно соединить последовательностью ребер $\{e_{t_q, t_{q+1}}, q = 0, \ldots, m-1, m \geqslant 1\}$, где $e_{t_0 t_1} = e_{i,t_1}$, а $e_{t_{m-1} t_m} = e_{t_{m-1} j}$, т. е. связать вершины v_i и v_j путем.

Определение 9.4. Максимальный связанный подграф $G' = (V', E')$ графа $G = (V, E)$ называется *компонентой*.

Пусть $G' = (V', E')$ и $G'' = (V'', E'')$ — две компоненты данного графа $G = (V, E)$. Тогда непосредственно из определений вытекает, что V' и V'', как подмножества в V, имеют пустое пересечение.

9.3.2. Алгоритм выделения компонент графа. В рамках общего подхода, изложенного в [83, с. 83—111], каждый метод классификации на графах опирается на процедуру выделения соответствующих F-максимальных подграфов. Более подробно рассмотрим это в п. 9.3.4. Здесь же опишем процедуру, лежащую в основе ряда известных методов, связанных с выделением компонент графа.

Исходя из некоторой модели класса, описываемой в терминах теории графов (см. п. 9.3.3 и 9.3.4), строится подграф G' графа близости $G(O) = (V, E)$ с тем же множеством вершин V, т. е. $G' = (V, E')$, где E' получается из E удалением ребер, не отвечающих модели класса. Затем применяется алгоритм выделения компонент $G'_1 = (V'_1, E'_1)$, ..., $G'_k = (V'_k, E'_k)$ графа $G' = (V, E')$ и тем самым находится разбиение множества V на подмножества $V_1, \ldots, V_k$, т. е. классификация множества объектов O.

Указанная процедура, примененная к **B**-графу близости, позволяет провести классификацию при наличии матрицы ограничений **B** на связи между объектами.

Определение 9.5. Подграф $G' = (V', E')$ некоторого графа $G = (V, E)$ называется G-полным подграфом (короче, G-подграфом), если любое ребро из G, соединяющее какие-либо вершины из G', принадлежит G', т. е. $E' = \{e_{ij} \in E: v_i, v_j \in V'\}$.

Непосредственно из определения получаем:

1) существует взаимно-однозначное соответствие между подмножествами множества вершин V и G-подграфами графа $G = (V, E)$;

2) каждая компонента G' графа G является его G-подграфом.

Опираясь на эти утверждения, опишем основной этап алгоритма.

Шаг А. Пусть $G^1 = (V^1, E^1)$-связанный подграф графа $G = (V, E)$ и $V^1 = \{v_{i_1}, \ldots, v_{i_m}\} \subset V = \{v_1, \ldots, v_n\}$, $m \leqslant n$.

Положим $V^2 = \{v_j \in V : e_{ii_t} \in E, \ t = 1, \ldots, m\}$, т. е. введем множество вершин в G, связанных с вершинами из G' хотя бы одним ребром. Тогда G-подграф $G^2 = (V^2, E^2)$ графа G, соответствующий множеству вершин V^2, будет связанным. Если $V^2 = V^1$, то получаем, что G^2 является компонентой, содержащей данный граф G^1. Если $V^2 \neq V^1$, то аналогично построим множество вершин V^3 и G-подграф $G^3 = (V^3, E^3)$, соответствующий множеству вершин V^3. Так как V — конечное множество, то последовательность множеств $V^1 \subset V^2 \subset V^3 \subset \ldots$ стабилизируется, т. е. $V^q = V^{q+1}$ для некоторого q и получим G-подграф $G^q = (V^q, E^q)$, являющийся компонентой графа G, содержащей данный граф G^1.

Замечание. Программную реализацию шага А можно провести, оперируя только векторами инцидентности вершин (см. определение 9.2).

Итак, на вход алгоритма выделения компонент подается связанный граф, например граф близости $(v_1, \varnothing)$ первой вершины. Применяя шаг А, получаем компоненту $G_1 = (V_1, E_1)$ графа $G = (V, E)$ и переходим к выделению компонент графа $(V \setminus V_1, E \setminus E_1)$ и т. д. до тех пор, пока не исчерпаем все исходное множество вершин V.

9.3.3. Алгоритмы классификации, использующие процедуру выделения компонент графа. Рассмотрим матрицу взаимных близостей $\rho = (\rho_{ij} = \rho(O_i, O_j), \ 1 \leqslant i, j \leqslant n)$, между классифицируемыми объектами $O_1, \ldots, O_n$. Без ограничения общности можно считать, что $0 \leqslant \rho_{ij} \leqslant 1$, причем $\rho_{ij} = 0$ тогда и только тогда, когда $i = j$. Пусть $0 = c_0 < c_1 < \ldots < c_m = 1$ — некоторая последовательность пороговых значений. Тогда определена последовательность подграфов графа близости $G(O) = (V, E) : G^0 \subseteq G^1 \subseteq \ldots \subseteq G^m$, где $G^t = (V, E^t)$ и $E^t = \{e_{ij} \in E : \rho_{ij} \leqslant c_t\}$. Заметим, что $G^0 = (V, \varnothing)$ и $G^m = G$. Граф G^t называется графом близости на уровне c_t.

Опишем сначала общую схему алгоритмов.

Параметром алгоритма является оператор U, ставящий в соответствие последовательности графов $G^0 \subseteq \ldots \subseteq G^m$ последовательность графов $\widehat{G}^0 \subseteq \ldots \subseteq \widehat{G}^m$, где $\widehat{G}^t = (V, \widehat{E}^t)$, и $\widehat{E}^t$ получается из E^t удалением некоторых связей в соответствии с выбранной моделью классов. Основные примеры операторов U рассмотрены в конце этого и в следую-

щем пункте. Ясно, что $\widehat{G}^0 = G^0$, а в тех случаях, когда нет априорных ограннчений на связи между объектамн, то н $\widehat{G}^m = G^m$. В общем случае граф $\widehat{G}^m$ представляет собой **B**-граф блнзости, где $\mathbf{B} = (b_{ij})$ — матрнца ограничений на связи между объектамн.

Схема алгорнтма

1. На вход подается граф G^0, множество его компонент задает разбиение совокупностн объектов $O_1, \ldots, O_n$ на одноточечные классы.

2. На вход t-го шага, $t \geqslant 1$, подается граф $\widehat{G}^{t-1}$, разбитый на компоненты $\widehat{G}_1^{t-1}, \ldots, \widehat{G}_{k_{t-1}}^{t-1}$, и граф $\widehat{G}^t$. Так как $\widehat{G}^{t-1} \subset \widehat{G}^t$, то каждая компонента графа $\widehat{G}^{t-1}$ является связанным подграфом графа $\widehat{G}^t$. Поэтому можно, начиная, например, с $\widehat{G}_1^{t-1}$, при помощи алгорнтма выделения компонент (см. п. 9.3.2) получить компоненту $\widehat{G}_1^t$ графа $\widehat{G}^t$ н перейти к выделению компонент графа $\widehat{G}^t \setminus \widehat{G}_1^t$. Так как компоненты графа не пересекаются, то компонента графа $\widehat{G}^{t-1}$ может входить только в одну компоненту графа $\widehat{G}^t$. Следовательно, выделение компонент в $\widehat{G}^t \setminus \widehat{G}_1^t$ можно начннать с одной из компонент графа $\widehat{G}^{t-1}$, не вошедших в $\widehat{G}_1^t$, и т. д. до тех пор, пока не будут выделены все компоненты графа $\widehat{G}^t$. Итогом t-го шага является разбиение графа $\widehat{G}^t$ на компоненты $\widehat{G}_1^t, \ldots, \widehat{G}_{k_t}^t$ и, следовательно, разбиение множества вершин V на непересекающнеся подмножества $V_1^t, \ldots, V_{k_t}^t$.

Соответствующая классифнкацня $(S_1^t, \ldots, S_{k_t}^t)$ множества объектов называется классификацней на уровне c_t.

Алгоритм заканчивает работу на m-м шаге.

В результате получаем последовательность вложенных разбиений $S^0 \subset S^1 \subset \ldots \subset S^m$, т. е. нерархию на множестве объектов O.

Рассмотрим теперь нанболее известные примеры алгоритмов, основанные на простых моделях классов и соответственно простых операторах U. Построение операторов U для более сложных моделей классов требует привлечения

результатов теории графов и изложено в следующем пункте.

Определение 9.6. Алгоритм классификации на графах с тождественным оператором U, т. е. $\widehat{G^t} = G^t$ для всех t называется *односвязывающим*.

В случае когда последовательность пороговых значений $0 = c_0 < c_1 < \ldots < c_m = 1$ совпадает с последовательностью рангов для последовательности ρ_{ij}, $1 \leqslant i, j \leqslant n$, иерархическая классификация при помощи односвязывающего алгоритма совпадает с иерархической классификацией, получаемой классической агломеративной процедурой по принципу «ближайшего соседа» (см. § 8.2). Алгоритм называется односвязывающим, так как он соответствует модели класса, в которой каждый объект из данного класса связан с остальными объектами из этого класса по крайней мере одной связью.

Определение 9.7. Алгоритм классификации на графах с оператором U, удаляющим в графах G^t ребра (связи), идущие к вершинам степени, меньше k, $k \geqslant 2$, называется *k-связывающим*.

Название алгоритма объясняется тем, что он соответствует модели класса, в которой представитель данного класса связан с остальными, по крайней мере, k связями.

Определение 9.8. При данном k, $k \geqslant 1$ максимальный связанный подграф G' графа G называется *k-связкой*, если степень каждой его вершины не меньше k.

Заметим, что каждая k-связка графа G^t является компонентой графа $\widehat{G^t}$ для оператора U из определения 9.7. Следовательно, k-связывающий алгоритм выделяет в исследуемом множестве объектов все k-связки на каждом уровне близости.

9.3.4. Метод послойной классификации. Общий подход к построению алгоритмов классификации на графах. Рассмотрим сначала, какие максимальные подграфы наряду с k-связкой используются для описания структур классов.

Определение 9.9. При данном k, $k \geqslant 1$, максимальный связанный подграф $G' = (V', E')$ графа G называется *k-компонентой*, если при любом разбиении множества вершин V' на непересекающиеся подмножества V'_1, V'_2 существует, по крайней мере, k связей (ребер) в E' между подмножествами V'_1 и V'_2.

Определение 9.10. Связанный подграф $G' = (V', E')$ графа G называется *k-связанным*, $k \geqslant 1$, если $|V'| \geqslant k + 1$, и удаление любых $(k - 1)$ вершин из V' оставля-

ет его связанным, т. е. любой подграф $G'' = (V'', E'') \subset G'$, у которого $|V''| > |V'| - k + 1$ является связанным.

Определение 9.11. При данном k, $k \geqslant 1$, максимальный k-связанный подграф $G' = (V', E') \subset G$ называется *k-блоком*.

Определение 9.12. *Кликой графа* G называется максимальный полный подграф $G' = (V', E')$. Клика графа, содержащая более k вершин, называется *k-кликой*.

Непосредственно из определений 9.8—9.12 следует:

1) для любого графа G каждая k-клика является подграфом некоторого k-блока, каждый k-блок является подграфом

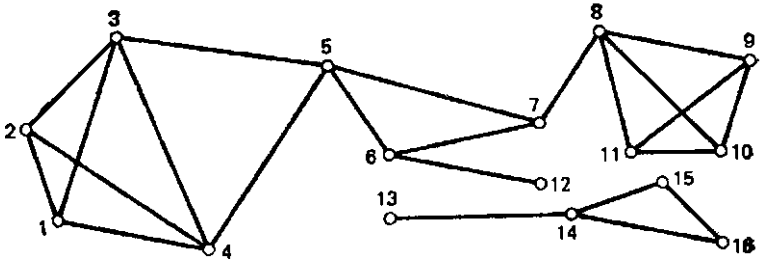


Рис. 9.1. Классификация на графе.
Компоненты {1, ..., 12}, {13, ..., 16};
2-связки: {1, .., 11}, {14, 15, 16};
2-компоненты {1, ..., 7}; {8, ..., 11}, {14, 15, 16};
2-блоки. {1, ..., 5}; {5, 6, 7}, {8, ..., 11}, {14, 15, 16};
2-клики: {1, 2, 3, 4}, {3, 4, 5}, {5, 6, 7}, {8, ..., 11}, {14, 15, 16}

некоторой k-компоненты, которая в свою очередь является подграфом некоторой k-связки (рис. 9.1, где $k = 2$);

2) любые две k-компоненты и, следовательно, любые две k-связки одного графа G имеют пустое пересечение;

3) пересечение любых двух k-блоков одного графа может содержать не более чем $(k - 1)$ общих вершин.

Для удобства изложения будем считать, что каждая вершина графа G, не вошедшая ни в один подграф, максимальный по отношению к выбранному свойству F, формально обладает этим свойством. Например, будем говорить, что вершина, не вошедшая ни в одну k-компоненту графа G, сама является его k-компонентой.

При таком соглашении можно сказать, что выделение в данном графе $G = (V, E)$ всех его подграфов $G_1 = (V_1, E_1)$, ..., $G_k = (V_k, E_k)$, максимальных по отношению к выбранному свойству F, задает покрытие множества вершин V подмножествами V_1, ..., V_k, $\bigcup_{q=1}^{k} V_q = V$.

В случае k-компонент и k-связок получается разбиение множества V на непересекающиеся подмножества, а в случае k-блоков — покрытие, каждые два множества которого имеют не более чем $(k-1)$ общую точку.

Теперь можно описать общую схему алгоритма классификации на графах как естественное обобщение схемы из п. 9.3.2.

Пусть, как и выше, $G^0 \subseteq G^1 \subseteq \ldots \subseteq G^m$ — последовательность подграфов графа G, где G^t — граф близости на уровне c_t. Параметром алгоритма является процедура A_F выделения в графе G^t всех его подграфов, максимальных по отношению к выбранному свойству F. Эти подграфы будем называть F-компонентами графа G. Роль F-компонент, в зависимости от F, играют k-компоненты, k-связки, k-блоки или k-клики.

Схема алгоритма

1. На вход подается граф G^0, множество его F-компонент задает покрытие совокупности объектов $\{O_1, \ldots, O_n\}$ одноточечными классами.

2. На вход t-го шага, $t \geqslant 1$, подается граф G^{t-1}, покрытый F-компонентами $G_1^{t-1}, \ldots, G_{k_{t-1}}^{t-1}$ и граф G^t. Так как $G^{t-1} \subset G^t$, то каждая F-компонента графа G^{t-1} является связным подграфом в G^t. На вход процедуры A_F подается граф G_1^{t-1}. При помощи нее он достраивается до F-компоненты G_1^t графа G^t. Затем среди F-компонент графа G^{t-1} берется та, которая не вошла в G_1^t, и при помощи той же процедуры достраивается до следующей F-компоненты графа G^t и так далее, до тех пор, пока не будут выделены все F-компоненты $G_1^t, \ldots, G_{k_t}^t$ графа G^t. Соответствующее покрытие $(S_1^t, \ldots, S_{k_t}^t)$ совокупности объектов называется F-классификацией ее на уровне c_t. Классы S_q^t, $q = 1, \ldots, k_t$, могут пересекаться, как в случае классификации k-блоками, но ни один класс не может целиком содержать другой.

Алгоритм заканчивает работу на $(m-1)$-шаге, так как G^m является графом близости совокупности объектов, т. е. полным графом. (Это в случае, когда нет ограничений на связи между объектами, а если матрица априорных ограничений имеется, то алгоритм заканчивает работу на m-м шаге.)

В результате получается последовательность вложенных покрытий $S^0 \subset S^1 \subset \ldots \subset S^m$, которая называется послойной F-классификацией. В тех случаях, когда для каж-

дого t покрытие S^t состоит из непересекающихся классов, то F-классификация является иерархической, т. е. задает иерархию на множестве объектов (см. определение 8.1).

Отмеченные выше взаимоотношения между F-компонентами для различных свойств F позволяют рассматривать в целом метод послойной классификации как последовательность уточняющих друг друга методов послойных F-классификаций. Классификация k-компонентами уточняет классификацию k-связками и сама в свою очередь уточняется классификацией k-блоками и т. п.

Эффективность алгоритмов послойных F-классификаций определяется эффективностью реализации процедуры A_F выделения в графе всех его F-компонент. Как следует из п.9.3.3, в случае k-связок существует достаточно эффективная процедура A_F благодаря тому, что каждая k-связка графа G^t является компонентой графа $\widehat{G^t}$, полученного из G^t удалением некоторых связей.

Непосредственные процедуры A_F для F-классификаций, уточняющих классификацию k-связками, очень трудоемки, и реализация их затруднена даже для относительно небольших совокупностей объектов. Но, привлекая результаты теории графов, удается существенно сократить наиболее трудоемкую часть процедур выделения F-компонент, связанную с перебором вариантов разбиения множества вершин.

Для построения алгоритмов классификации k-компонентами и k-блоками большое значение имеют следующие результаты, полученные К. Менгером [83, с. 102—106].

Т е о р е м а 9.1. а) Минимальное число связей, удаление которых разъединяет две какие-либо вершины графа, равно максимальному числу не содержащих общих связей путей между этими двумя вершинами; б) минимальное число вершин, удаление которых разъединяет какие-либо две несмежные (не связанные непосредственно) вершины графа, равно максимальному числу непересекающихся (за исключением концов) путей между этими двумя вершинами.

С л е д с т в и е 9.1. Каждая пара v_i, v_j различных вершин k-компоненты G' графа G соединена k путями подграфа G', не содержащими общих связей, причем подграф G' — максимальный подграф с этим свойством.

С л е д с т в и е 9.2. Каждая пара v_i, v_j различных вершин k-блока G' графа $G=(U, E)$ соединена k путями подграфа G', не содержащими общих точек (за исключением концов), причем G' — максимальный подграф с этим свойством, если $|U| \geqslant k+1$.

Опишем, например, как на основе следствия 9.1 построить процедуру A_F выделения k-компонент графа G^t.

Схема алгоритма

1. Применяя процедуру выделения k-связок из п. 9.3.3, разобьем граф G^t на k-связки $G_1^t, \ldots, G_{m_t}^t$.

2. Пусть G_q^t — некоторая k-связка. Выделим все одноточечные k-компоненты. Согласно следствию 9.1 вершина v_i графа G_q^t является одноточечной компонентой тогда и только тогда, когда существует хотя бы одна другая вершина v_j этого графа, такая, что v_i и v_j соединяет меньше, чем k путей, не содержащих общих связей.

3. Пусть $\widetilde{G}_q^t$ — подграф k-связки G_q^t, у которой удалены все одноточечные компоненты. Так как любые две k-компоненты одного графа не пересекаются, то $\widetilde{G}_q^t$ представляет собой граф, связные компоненты которого являются k-компонентами графа $\widetilde{G}_q^t$. (Фактически здесь описана конструкция оператора удаления U (см. п. 9.3.3) для получения k-компонент.) Применяя теперь алгоритм выделения компонент графа $\widetilde{G}_q^t$, получаем, наконец, набор k-компонент графа G_q^t.

Применяя шаги 2 и 3 последовательно к k-связкам G_q^t, $q = 1, \ldots, m_t$, получаем набор всех k-компонент графа G^t.

ВЫВОДЫ

1. Алгоритмы разделения смесей легко модифицируются для работы при наличии неполных обучающих выборок. Такой тип задания априорной информации весьма эффективен — известны примеры, когда использование ОВ объема 10 % исходной совокупности резко улучшало результаты классификации.

2. Описаны методы и алгоритмы классификации в ситуации, когда исследователь желает получить разбиение объектов на классы, согласованное с ограничениями на связи между объектами.

3. Описаны методы и алгоритмы классификации, основанные на представлении исходной информации о классифицируемых объектах в виде последовательности подграфов

$G^0 \subseteq \ldots \subseteq G^m$ графа **близости** G, где G^t — граф близости на уровне t-го порога.

4. Каждый метод классификации опирается на модель класса в виде максимального подграфа (F-компоненты) одиого из графов близости G^t, $t = 0, \ldots, m$, где F — свойство класса. В качестве моделей классов рассматриваются k-связки, k-компоненты, k-блоки и k-клики.

5. Описана общая схема алгоритмов, задающих последовательность покрытий $S^0 \subset \ldots \subset S^m$ совокупности объектов $O_1, \ldots, O_n$, где S^t — покрытие, образованное всеми F-компонентами графа G^t. Рассмотрены важнейшие примеры алгоритмов и взаимосвязи между получаемыми с нх помощью классификациями.

Глава 10. ТЕОРИЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

10.1. Математическая модель алгоритма автоматической классификации (ААК)

Как и выше, предполагаем, что исходиая информация об исследуемой совокупиости (выборке) объектов $O = (O_1, \ldots, O_n)$ представлена либо в форме матрицы «объект — свойство» $\mathbf{X} = (x_i^{(j)})$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, либо матрицы взаимных близостей $\rho = (\rho_{ij})$, $1 \leqslant i, j \leqslant n$. В случае числовых признаков отождествляем объект O_i с точкой $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})$ евклидова пространства R^p.

Начнем с развернутой характеристики компонеит математической модели алгоритма автоматической классификации (АК) [10, 41].

10.1.1. Пространство состояний S_O. Эта компонента описывает допустнмые классификации, возникающие на шагах алгорнтма. Так, если результатом m-го шага является разбиение $S = (S_1, \ldots, S_k)$ совокупности O на k непересекающихся классов, то S_O — это пространство всех разбиений множества из n элементов на фнксированное или нефиксированное чнсло классов в зависимости от того, не зависит или зависнт k от номера шага. В алгоритмах метода динамнческнх сгущений роль S_O играет пространство покрытнй (см. § 7.4).

В рассматриваемой модели каждый элемент $S \in S_O$ отождествляется с некоторым отображением выборки O в множество Z — множество значений классификаций. Состав и структура Z определяют тип решаемой задачи, т. е., по

существу, дают средство ответить на вопрос. какую задачу классификации предполагается решить? Например, когда $Z = \{1, \ldots, k\}$ — список номеров классов, то $S_O = \{s : O \to Z\}$ совпадает с множеством разбиений выборки O на k непересекающихся классов

Определение 10.1. Будем считать, что классификация s: $O \to Z$ выборки O согласована с информацией, выражаемой функцией $\varphi : O \to 2^Z$ (обозначение $s \in \varphi$), если $s(O_i) \in \varphi(O_i)$ для любого i. В этом случае пространство допустимых классификаций будет иметь вид: $S_O = \{s : O \to Z : s \in \varphi\}$.

Таким образом, чтобы задать пространство состояний S_O, необходимо формализовать задачу, т. е. указать Z и сформулировать условия, выделяющие S_O в множестве всех отображений из O в Z.

10.1.2. Пространство описаний L. Эта компонента связана с выбором средств информативного описания классов для достижения цели классификации. Например, в алгоритмах метода динамических сгущений роль L играет пространство представительств (см. п. 7.4.1).

В рассматриваемой модели каждый элемент $l \in L$ отождествляется с некоторым отображением Z в Y_O, где Y_O — множество значений, в терминах которых выражается результат классификации.

Пример 10.1. Пусть $O = \{X_1, \ldots, X_n\} \subset R^p$ и $Z = \{1, \ldots, k\}$. Тогда, если описывать каждый класс $S \subset O$ его средним $\frac{1}{|S|} \sum_{X \in S} X \in R^p$, то $Y_O = R^p$ и

$$L = \{Z \to Y_O\} = R^p \times \ldots \times R^p = (R^p)^k.$$

Пример 10.2. Пусть имеется набор опорных точек $Y_1^0, \ldots, Y_k^0$, $Y_q^0 \in R^p$, причем известно, что ядро q-го класса Y_q, $Y_q \in R^p$, должно находиться в окрестности радиуса r_q точки Y_q^0, причем случай $r_q = 0$ для некоторых q не исключается, то

$$L = \{(Y_1, \ldots, Y_k),\ Y_q \in R^p,\ \|Y_q - Y_q^0\| \leqslant r_q,\ q = 1, \ldots, k\}.$$

Роль опорных точек $\{Y_q^0\}$ могут играть как реальные представители классов, так и точки, построенные на основе дополнительных сведений — формальные представители классов.

Пример 10.3. Пусть предполагается описывать класс $S \subset O$ набором из m его представителей. Тогда Y_O — совокупность всех m — элементных подмножеств в O.

Пространство Y_O может иметь совершенно нную природу, чем пространство измеряемых признаков. Так, если в качестве ядра класса брать проходящее через центр класса линейное пространство главных факторов этого класса, то Y_O является многообразием линейных подпространств в в R^p. Более того, само Y_O может быть составлено из пространств различной природы соответственно предположениям о природе различных классов (см. данное ииже описанне модели алгоритма Форель).

Таким образом, чтобы задать пространство опнсаний L, необходимо выбрать, в каких терминах выражать результат классификации, т. е. указать пространство информативиого описания классов и сформулировать условия, выделяющие L в множестве всех отображеиий из Z в Y_O.

10.1.3. Множество порций P, в которых выборка поступает на классификацию и генератор порций G. Эта компонента вводится для того, чтобы в рамках модели можно было рассматривать алгоритмы как параллельиого ($P = \{O_i\}$ состоит из одного элемента, символизирующего всю выборку), так и последовательного типа (например, если на классификацию поступает по одному объекту, то $P = O$).

Генератор порций G описывает способ получения из выборки O порций объектов для проведения очередного шага алгоритма. В общем случае G представляет собой оператор, значение которого на $(m+1)$-м шаге зависит от результата классификации s_m и ее описания l_m, полученных на предыдущем шаге. Например, в алгорнтмах, использующих пороговые значения (при классификации выборок большого объема), генератор G из всей выборки отбирает только те элементы, которые отстоят от ядер классов, полученных на m-м шаге, не более чем на пороговое значение c_{m+1}.

10.1.4. Классификатор K. Эта компонента представляет собой оператор из $S_O \times L \times P$ в S_O, называемый классификатором, поскольку он определяет, что нужно сделать с имеющимися средствами Y_O для данного типа Z задачи АК, чтобы из предыдущего состояния s_m выборкн O с опнсаннем l_m перейти в другое состояние l_{m+1} при поступлении очередной порции $p_{m+1} \subset O$, т. е. провести переклассификацию. Частным случаем оператора K является функция иазначения $g : L \to S_O$ (см. § 7.4). Обычно оператор K строят нсходя из фуикционала F, оценнвающего качество класснфикации согласно априорному представлению нсследователя о «хорошей» классификацни.

Пример 10.4. Опишем, как строится оператор K в нсторически одном из первых и наиболее общем алгорнтме разбиения на k классов методом локальной оптнмизации.

Пусть $s : O \to Z = [1, \ldots, k]$ — некоторая классификация и $X \in O$. Обозначим через $s(q, X)$ новую классификацию: $s(q, X)(X') = s(X')$, если $X \neq X'$ и $s(q, X)(X) = q$. Фиксируем критерий качества классификации $F : S_O \to R^1$ и определим оператор $K : S_O \times P \to S_O$ для алгоритмов последовательного типа, т. е. когда $P = O$, формулой

$$K(s, X) = s, \text{ если } F(s) = \min_q F(s(q, X)); \tag{10.1}$$

$$K(s, X) = s(q, X), \text{ где } q \neq s(X) \text{ и } q = \arg\min_{q'} F(s(q', X)).$$

Формула 10.1 для классификатора K не налагает практически никаких ограничений на вид функционала F, но в общем случае приводит к довольно медленной процедуре классификации, реализующей минимизацию функционала F способом перебора на дискретном множестве S_O.

Пример 10.5. Предположим, что $F(s)$ имеет вид $\sum_{q=1}^{k} F_q(O(q))$, где $O(q) = s^{-1}(q)$ и $F_q(\cdot)$ — критерий однородности q-го класса, представляющий собой функцию на множестве всех подмножеств из O, такую, что $F_q(S_1) \leqslant F_q(S_2)$, как только $S_1 \subset S_2$. Тогда классификатор K можно задать следующим образом. Пусть

$$\Delta(q, X) = F_q(O(q) \cup X) - F_q((O(q) \cup X) \setminus X). \tag{10.2}$$

Положим

$$K(s, X) = s, \text{ если } \Delta(s(X), X) = \min_q \Delta(q, X);$$
$$K(s, X) = s(q, X), \text{ если } q \neq s(X) \text{ и } q = \arg\min_{q'} \Delta(q', X). \tag{10.3}$$

Опишем алгоритм равномерного распределения объектов по классам, в котором классификатор K действует по формуле (10.3). В этом алгоритме критерием качества разбиения выборки O на классы $O(1), \ldots, O(k)$ является сумма попарных внутриклассовых расстояний между объектами (см. п. 5.4.1), т. е. $F(s) = \sum_{q=1}^{k} F_q(O(q))$, где

$$F_q(O(q)) = \sum_{X_i, X_j \in O(q)} \|X_i - X_j\|^2. \tag{10.4}$$

Пусть $d(O(q))$ — статистический разброс класса $O(q)$, Y_q — центр этого класса и n_q — число элементов в нем. Заметим, что $F_q(O(q)) = 2n_q d(O(q))$.

Пусть классификация s относит данный объект $X \in O$ к классу $O(t)$.

Положим $O'(t) = O(t) \setminus X$. Непосредственно из формул (10.2) и (10.4) получаем:

$\Delta(q, X) = 2(d(O(q)) + n_q \| X - Y_q \|^2)$, если $q \neq t$ и

$\Delta(t, X) = 2(d(O'(t)) + (n_t - 1) \| X - Y_t' \|^2)$,

где Y_t' — центр класса $O'(t)$. Таким образом, процедуру перехода к новой классификации $K(s, X)$ можно в этом случае описать следующим образом.

Дана выборка $\mathbf{O}' = O \setminus X$, разбитая на классы $O'(q)$, $q = 1, \ldots, k$, где $O'(q) = O(q)$, если $q \neq t$. Поступает на классификацию объект X, не принадлежащий выборке O'. Рассчитываем для каждого класса $O'(q)$ его центр Y_q', статистический разброс $d(O'(q))$ и число элементов n_q'. Меру близости $\rho(X, O'(q))$ между элементом X и классом $O'(q)$ задаем формулой:

$$\rho(X, O'(q)) = d(O'(q)) + n_q' \| X - Y_q' \|^2. \qquad (10.5)$$

Тогда классификация $K(s, X)$ относит элемент X к ближайшему классу в смысле меры близости (10.5).

Классификатор K (10.3) применим и в случае, когда имеется дополнительная информация о допустимой принадлежности элементов к классам, т. е. имеется функция $\varphi: O \to 2^Z$ (см. определение 10.1). Тогда, например в алгоритме равномерного распределения объектов классификатор $K(s, X)$ отнесет элемент X к ближайшему из классов $\mathbf{O}'(q)$, где

$q \in \varphi(X) \subset Z = [1, \ldots, k]$.

Рассмотренные выше классификаторы (10.1) и (10.3) действуют только в процедурах последовательного типа[1]. Для построения классификатора в процедурах параллельного типа обычно используются функционалы $F: S_O \times L \to \to R^1$ вида

$$F(s, l) = \sum_{q=1}^{k} \sum_{s(x)=q} F_q(X, l(q)),$$

где $F_q(X, l(q))$ описывает, насколько объект X близок ядру $l(q)$. В этом случае классификатор K можно взять в виде:

$$K(s, l, p)(X) = \begin{cases} q = \arg\min F_q(X, l(q)), & X \in p, \\ s(X), & X \overline{\in} p. \end{cases} \qquad (10.6)$$

Если функционал $F_q(X, Y)$ не зависит от номера класса q,

[1] Точнее, только в процедурах последовательного типа можно гарантировать убывание глобального критерия $F(s)$ при переходе к новой классификации.

то классификатор (10.6) совпадает с функцией назначения $g: L \to S_O$ (см. § 7.4).

10.1.5. Дескриптор D. Эта компонента представляет собой оператор из $S_O \times L$ в L, называемый *дескриптором*, поскольку с его помощью на основании предыдущего описания l_m выборки O и полученной классификации s_{m+1} вырабатывается новое, более оптимальное описание. Частным случаем оператора D является функция представительства $S_O \to L$. В алгоритмах АК, основанных на описании классов ядрами, оператор D строят исходя из функционала $F: S \times L \to R^1$ вида

$$F(s, l) = \sum_{q=1}^{k} F_q(O(q), l(q)),$$

по формуле:

$$D(s, l)(q) = \arg\min_{Y \in Y_O} F_q(O(q), Y). \tag{10.7}$$

Таким образом, для вычисления значений $D(s, l)(q)$ дескриптора D необходимо применить некоторую процедуру минимизации функционала $F_q(O(q), Y)$ на Y_O. Наибольшее распространение получили алгоритмы АК, в которых в качестве $F_q(O(q), Y)$ используется статистический разброс класса $O(q)$ относительно $Y \in Y_O$, так как в этом случае удается получить явное решение задачи минимизации и заменить формулу (10.7) явной формулой расчета ядра q-го класса по элементам этого класса.

Пример 10.6. В алгоритме k-средних имеем $Y_O = R^p$ и $F_q(O(q), y) = \sum_{X \in O(q)} ||X - Y||^2$. Поэтому

$$D(s, l)(q) = \frac{1}{|O(q)|} \sum_{X \in O(q)} X.$$

Пример 10.7. Опишем алгоритм типологического главного фактора [106]. Пусть $O = \{X_1, \ldots, X_n\} \subset R^p$ и $Z = \{1, \ldots, k\}$. Ядро класса Y описывается парой (Y, v), $Y \in R^p$, $v \in R^p$, $||v|| = 1$, а мерой близости от точки X до ядра $\tilde{Y} = (Y, v)$ считается квадрат расстояния от точки X до прямой, проходящей через точку Y с направляющим вектором v, т. е.

$$F_q(X, \tilde{Y}) = \|(X - Y) - v'(X - Y)v\|^2. \tag{10.8}$$

Допустим, что на m-м шаге алгоритма совокупность O разбита на k классов $O(1), \ldots, O(k)$. Статистический разброс

класса $O(q)$ относительно $\widetilde{Y} = (Y, v)$ вычисляется по формуле:

$$F_q(O(q), \widetilde{Y}) = \sum_{X \in O(q)} \|(X-Y) - v'(X-Y)v\|^2. \qquad (10.9)$$

Минимизируя функционал (10.9) на пространстве $Y_O = \{\widetilde{Y} = (Y, v),\ Y \in R^p,\ v \in R^p,\ \|v\| = 1\}$, получаем, что $D(s, l), (q) = (Y(q), v(q))$, где $Y(q)$ — центр класса $O(q)$, $v(q)$ — собственный вектор с наибольшим собственным значением ковариационной матрицы класса $O(q)$, т. е. главная компонента этого класса. На $(m+1)$-м шаге этого алгоритма применяется классификатор (10.6), т. е. стронтся минимальное дистанционное разбиение, порождаемое набором ядер $\{(Y(1), v(1),) \ldots, (Y(k), v(k))\}$ для меры близостн (10.8). Алгоритм останавливается, если новое разбиение имеет тот же набор ядер, что предыдущее.

10.1.6. Основные понятия и определения, используемые при исследовании математической модели АК. Опираясь на материал п. 10.1.1 — 10.1.5, естественно дать следующие определения.

Определение 10.2. *Моделью алгоритма* АК называется набор его компонент $(S_O, L, P; K, D, G)$, наделенных опнсанной выше структурой.

Определение 10.3. *Движением алгоритма*, отвечающим начальным данным $s_0 \in S_O$, $l_0 \in L$ и $p_0 \in P$, называется последовательность $(s_0, l_0), (s_1, l_1), \ldots, (s_m, l_m), \ldots$, где $s_{m+1} = K(s_m, l_m, p_m)$, $l_{m+1} = D(s_{m+1}, l_m)$, $p_{m+1} = G(p_m, s_{m+1}, l_{m+1}, m+1)$.

Определение 10.4. Алгоритм АК называется *сходящимся*, если для его движения $(s_0, l_0), \ldots, (s_m, l_m), \ldots$ последовательность описаний $l_0, \ldots, l_m, \ldots$ сходится в метрике пространств L к некоторому предельному описанию l_*.

Алгоритм АК называется *стабилизирующимся*, если существует номер m_0, такой, что $l_m = l_{m_0}$ для всех $m \geqslant m_0$.

Определение 10.5. Функционал $F: S_O \times L \to R^1$ называется интерпретирующим для модели ААК, если

$$F(s_m, l_m) \geqslant F(s_{m+1}, l_m); \qquad (10.10)$$

$$F(s_{m+1}, l_m) \geqslant F(s_{m+1}, l_{m+1}) \qquad (10.11)$$

для всех m, начиная с некоторого m_0.

Таким образом, если функционал F рассматривать как меру потерь при задании выборки O в состоянии (классификации) s ее описанием l, то неравенства (10.10) и (10.11) служат объяснением, почему от состояния s_m при фиксирован-

ном l_m переходим к состоянию $s_{m+1} = K(s_m, l_m, p_m)$ и от описания l_m выборки O при фиксированном s_{m+1} переходим к $l_{m+1} = D(s_{m+1}, l_m)$. Для данной модели алгоритма АК, очевидно, может существовать много интерпретирующих функционалов, причем роль их при исследовании данного алгоритма может быть различной

Пример 10.8. Рассмотрим алгоритм Форель (см. п. 7 2.1), выделяющий в выборке $O = \{X_1, \ldots, X_n\} \subset R^p$ несмещенную подвыборку $O_* = \{X \in O: |X - Y_*| \leqslant r\}$, где r — параметр алгоритма, а Y_* — центр класса O_*, который находится как результат стабилизации последовательности описаний $Y_0, \ldots, Y_m, \ldots$

При исследовании этого алгоритма используются два интерпретирующих функционала:

$$F_1(O_m, Y_m) = \sum_{X \in O_m} \|X - Y_m\|^2 + r^2(n - n_m);$$

$$F_2(O_m, Y_m) = \sum_{X \in O_m} (\|X - Y_m\|^2 - r^2) =$$

$$= -r^2 \sum_{X \in O} \left(1 - \frac{\|X - Y_m\|^2}{r^2}\right)_+,$$

где $(\cdot)_+ = (\cdot)$, если $(\cdot) > 0$, и $(\cdot)_+ = 0$, если $(\cdot) < 0$. Здесь O_m — класс, выделенный на m-м шаге алгоритма, $Y_m \in R^p$ — описание класса (его центр) на этом шаге и $n_m = |O_m|$.

Используя F_1, получаем, что стабилизируемость движения алгоритма Форель является следствием легко доказываемой стабилизируемости движения алгоритма 2-средних (см. [41, 42]). В терминах F_2 доказать стабилизируемость труднее [27], но зато, используя этот функционал, получаем, что предельное описание Y_* является точкой локального максимума оценки плотности распределения случайного вектора по выборке O (см. § 7.6).

10.2. Базисная модель алгоритма АК, основанного на описании классов ядрами

Пусть $F_q(X, Y)$ — некоторая мера близости между объектом X и точкой Y пространства Y_q, называемого пространством ядер q-го класса. Для класса $O(q) \subset O$ и точки $Y \in Y_q$ определим меру близости $F_q(O(q), Y)$ формулой

$$F_q(O(q), Y) = \sum_{X \in O(q)} F_q(X, Y). \qquad (10.12)$$

Определение 10.6. Математическая модель АК, в которой $S_O \subset \{O \to Z\}$, где $Z = \{1, \ldots, k\}$ — список имен классов, и классификатор задается формулой

$$K(s, l, p)(X) = \begin{cases} q = \arg\min F_q(X, l(q')), & X \in p \subset O, \\ s(X), & X \overline{\in} p, \end{cases} \quad (10.13)$$

а дескриптор — формулой

$$D(s, l)(q) = \arg \min_{Y \in Y_q} F_q(O(q), Y), \quad (10.14)$$

где $F_q(O(q), Y)$ имеет вид (10.12), называется *базисной моделью алгоритма АК*, основанного на описании классов ядрами.

Далее для краткости эту модель алгоритма будем называть *базисной моделью ядерного алгоритма.*

Опишем ядерный алгоритм, в модели которого мера близости $F_q(X, Y)$ между объектом X и ядром q-го класса Y зависит от дополнительной информации об этом классе.

Пример 10.9. Пусть $O = \{X_1, \ldots, X_n\} \subset R^p$ — классифицируемая совокупность объектов и $V = \{V_1, \ldots, V_k\} \subset R^p$ — обучающая выборка, где $V_q = \{Y_{q1}^0, \ldots, Y_{qn_q}^0\}$ — представители q-го класса, причем $\sum_{q=1}^{k} n_q \ll n$ и не исключается, что множества V_q для некоторых q пустые.

Рассмотрим задачу разбиения выборки O на k классов. Так как обучающая выборка V мала, то применить обычные процедуры классификации при наличии обучающих выборок не представляется возможным. Для решения этой задачи можно рекомендовать алгоритм АК, описываемый базисной моделью ядерного алгоритма с

$$F_q(X, Y) = \sum_{t=1}^{T_q} \|Y_{qt} - X\|^2 + \sum_{j=1}^{n_q} \|Y_{qj}^0 - X\|^2, \quad (10.15)$$

где $Y = (Y_{q1}, \ldots, Y_{qT_q})$ — ядро q-го класса, составленное из точек пространств R^p, а T_q — максимально возможное число вводимых эталонов в q-м классе, $q = 1, \ldots, k$.

Любой алгоритм АК, описываемый базисной моделью ядерного алгоритма, допускает две важные модификации.

1. *Введение класса «джокер»* (так называемый класс «не знаю», «отказ» и т. п.). Обычно класс «джокер» определяется следующим образом Задается порог δ, и если для классифицируемого объекта мера близости $F_q(X, Y_q)$ превосходит δ для всех q, $1 \leqslant q \leqslant k$, где Y_q — ядро q-го класса на данном шаге алгоритма, то оператор K отказывается от клас-

сификации и относит элемент X к символическому эталону $*$ класса «джокер». На следующих шагах алгоритма, когда набор ядер $(Y_1, \ldots, Y_k)$ сменится, этот элемент может быть уже отнесен к некоторому классу. Для включения класса «джокер» в общую схему базисного ядерного алгоритма достаточно присвоить этому классу номер $(k+1)$ и положить $F_{q+1}(X, Y_{k+1}) = \delta$ для всех $X \in O$. Тогда $F_{q+1}(O(k+1), Y_{k+1}) = n_{k+1}\delta$, где $n_{k+1} = |O(k+1)|$ — число элементов в $(k+1)$-м классе, т. е. число элементов, которые на данном шаге не классифицированы.

2. Следующая модификация применяется тогда, когда имеется *дополнительная информация о допустимой принадлежности элементов к классам*, т. е. $\varphi : O \to 2^Z$. В этом случае классификатор задается формулой

$$K(s, l, p)(X) = \begin{cases} q = \arg \min\limits_{q' \in \varphi(X)} F_{q'}(X, l(q')), X \in p. \\ s(X), \ X \overline{\in} p \end{cases} \quad (10.16)$$

10.3. Иерархическая структура многообразия алгоритмов АК

В данной модели алгоритма АК одни и те же множества S_O и L могут быть получены для различных конкретизаций множеств Z и Y_O соответственно. Таким образом, в рамках данной модели алгоритма АК элементы Z, Y_O ее структуры являются варьируемыми параметрами. Покажем, как эти параметры позволяют восстанавливать иерархическую структуру многообразия алгоритмов.

10.3.1. Модель алгоритма k-средних параллельного типа. Пусть $\mathbf{O} = \{X_1, \ldots, X_n\} \subset R^p$. Для простоты будем считать, что p-мерное евклидово пространство R^p наделено стандартной метрикой: $||X - Y||^2 = \sum\limits_{i=1}^{p} (x^i - y^i)^2$, где $(x^1, \ldots, x^p)$ и $(y^1, \ldots, y^p)$ — координаты векторов X и Y соответственно.

В качестве первой конкретизацин множества Z возьмем список номеров классов $Z_1 = \{1, \ldots, k\}$. Положнм $S_O = \{s: O \to Z_1\}$ и будем отождествлять каждое отображение $s : O \to Z_1$ с разбиением множества O на классы $(O(1), \ldots, O(k))$, где $O(q) = s^{-1}(q)$. Имеем: $\bigcup\limits_{q=1}^{k} O(q) = O$ и $O(q) \cap \cap O(q') = \varnothing$, $q \neq q'$. Каждый класс будем описывать его средним, т. е. в качестве Y_O возьмем само пространство R^p и положим $L = \{l : Z_1 \to Y_O\} = R^p \times \ldots \times R^p$,

ким образом $l = (Y_1, \ldots, Y_k)$ и потому в L можно ввести метрику следующим образом: $\|l_1 - l_2\|^2 = \sum_{q=1}^{k} \|Y_{1q} - Y_{2q}\|^2$, где $l_1 = (Y_{11}, \ldots, Y_{1q})$ и $l_2 = (Y_{21}, \ldots, Y_{2q})$.

Так как строим модель алгоритма параллельного типа, то в качестве P должны взять одноэлементное множество (выборка O вся участвует в классификации на каждом шаге), поэтому в дальнейших формулах компоненты P и G модели можно не учитывать. Операторы K и D возьмем, следуя базисной модели ядерного алгоритма с $F_q(X, Y) = \|X - Y\|^2$. Детальное описание этих операторов дано в п. 7.2.1. Непосредственно из конструкции этих операторов следует, что функционал

$$F(s, l) = \sum_{q=1}^{k} \sum_{X \in O(q)} \|X - Y(q)\|^2, \tag{10.17}$$

где $s = (O(1), \ldots, O(k))$, $l = (Y(1), \ldots, Y(k))$, является интерпретирующим для этой модели (см. §. 10.4). Итак, модель алгоритма k-средних параллельного типа описана. Начнем модификацию ее параметров.

10.3.2. Модель алгоритма $(k - r)$-средних. Эта модель связана с введением класса «джокер» (см. § 10.2). Задается «порог отказа» r, и если значение меры сходства объекта X с каждым из k ядер классов превосходит r, то такой объект относится к символическому классу $*$.

Итак, $Z_1(*) = \{1, \ldots, k, k+1\}$.

В этом случае S_O точно такое же, как в алгоритме $(k+1)$-средних.

Далее, $Y_O = Y_1 \cup Y_2 = R^p \cup Y(*)$ — объединение пространства R^p и изолированной точки $Y(*)$. В качестве L возьмем часть множества отображений $Z_1(*)$ в Y_O, состоящую из всех отображений, переводящих символ $(k+1)$ в $Y(*)$. Тогда каждая точка $l \in L$ будет иметь вид $(Y(1), \ldots, Y(k), Y(*))$.

Операторы K и D зададим формулами (10.13) и (10.14) соответственно с

$$F_q(X, Y(q)) = \|X - Y(q)\|^2, \quad q = 1, \ldots, k;$$
$$F_{k+1}(X, Y(*)) = r^2.$$

Имеем:

$$O(k+1) = O(*) = \{X \in O : \min_{1 \leqslant q \leqslant k} \|X - Y(q)\|^2 > r^2\};$$

$$F_{k+1}(O(*), Y(*)) = |O(*)|\, r^2.$$

Таким образом описана модель алгоритма $(k-r)$-средних и тем самым определено его движение. Теперь непосредственной проверкой можно убедиться, что функционал

$$F(s, l)=\sum_{q=1}^{k}\sum_{X\in O(q)}\|X-Y(q)\|^2+|O(*)|r^2 \qquad (10.18)$$

является интерпретирующим для этой модели.

Можно подвести первые итоги.

Получение модели алгоритма $(k-r)$-средних иллюстрирует подъем снизу вверх в описываемой иерархии. Действительно, отправляясь от модели 10.3.1, получаем данную модель с дополнительным параметром r. Утверждение, что модель 10.3.2 лежит выше по иерархии, чем модель 10.3.1, обосновывается тем, что, отправляясь от модели 10.3.2 и устремляя $r\to\infty$, очевидно, получим модель 10.3.1.

10.3.3. Модель алгоритма Форель. Для получения этой модели (см. п. 7.2.1) достаточно, отправляясь от модели 10.3.2, спуститься вниз по иерархии, полагая $k=1$. Этот результат иллюстрирует способность данной иерархической структуры устанавливать связи между известными алгоритмами, которые появились для решения разных задач классификации. Между моделями 10.3.1 и данной, лежащими на одном уровне иерархии, устанавливается связь через модель 10.3.2, лежащую выше.

Применение алгоритма Форель опирается на гипотезу: выборка O представляет собой объединение $O(1)\cup O(2)$, где $O(1)$ — компактный кластер, ядро которого совпадает с его геометрическим центром, а точки из $O(2)$ лежат на достаточном удалении от кластера $O(1)$. Следующая модель описывает алгоритмы, опирающиеся на аналогичную гипотезу в предположении, что ядро компактного кластера $O(1)$ совпадает с взвешенным центром.

10.3.4. Модель алгоритма выделения размытого кластера. Параметром модели является невозрастающая функция $\gamma: R^1\to[0, 1]$, такая, что $\gamma(0)=1$ и $\gamma(t)=0$, $t\geqslant t_0$. В этой модели $Z=[0, 1]$ и пространство допустимых классификаций имеет вид $S_O=\{s: O\to[0, 1]: s(x)=\gamma(\|X-Y\|)$ для некоторого $Y\in R^p\}$. Согласно терминологии теории размытых множеств (см. п. 7.5.1) каждая такая классификация $s\in S_O$ задает размытый класс $O(1)$ в O. Класс $O(1)$ описывается точкой из R^p (формальным элементом, как и в алгоритме Форель), т. е. $Y_O=R^p$ и $L=R^p$. Разброс класса $O(1)$ относительно ядра $Y\in R^p$ задается формулой $F(O(1), Y)=\sum_{X\in O}\gamma(\|X-Y\|)\|X-Y\|^2$.

Для завершения описания модели осталось указать операторы K и D. Пусть на m-м шаге алгоритма получен размытый класс O_m (1) и его ядро Y_m. Тогда

$$K(s_m, Y_m)(X) = s_{m+1} = \gamma(\|X - Y_m\|);$$

$$D(s_{m+1}, Y_m) = Y_{m+1} = \arg\min_{Y \in R^p} \sum_{X \in O} \gamma(\|X - Y_m\|) \|X - Y\|^2,$$

т. е.

$$Y_{m+1} = \frac{\sum_{X \in O} \gamma(\|X - Y_m\|) X}{\sum_{X \in O} \gamma(\|X - Y_m\|)}.$$

Здесь s_m — описание класса O_m. Для получения из данного алгоритма алгоритма 10.3.3 надо спуститься вниз по иерархии, положив

$$\gamma(t) = \begin{cases} 1, & |t| \leqslant r; \\ 0, & |t| > r. \end{cases}$$

10.3.5. Модель алгоритма Δ(k)-средних. Опишем модель алгоритма, получение которой дает пример еще одного способа подъема снизу вверх в исследуемой иерархии. Для этого покажем сначала, как можно изменить параметры модели 10.3.1, совершенно не меняя ее основные компоненты, а значит, и движение алгоритма.

Обозначим через $\Delta(k)$ стандартный $(k-1)$-мерный симплекс, т. е. множество точек $\{(z^1, \ldots, z^k), z^q \geqslant 0, \Sigma z^q = 1\}$. Пронумеруем вершины $\Delta_1, \ldots, \Delta_k$ симплекса $\Delta(k)$ так, чтобы вершина Δ_q имела координаты $(0, \ldots, 0, 1, 0, \ldots, 0)$, где 1 стоит на q-м месте. Теперь в качестве второй конкретизации множества Z в модели 10.3.1 возьмем симплекс $\Delta(k)$. Тогда S_O из этой модели можно отождествить с частью множества отображений из O в $\Delta(k)$, состоящего из всех отображений, переводящих O в подмножество вершин $\{\Delta_1, \ldots, \Delta_k\} \subset \Delta(k)$.

Каждую точку $Z \in \Delta(k)$ можно однозначно записать в виде $Z = \sum_{q=1}^{k} z^q \Delta_q$, где $z^q \geqslant 0$, $\Sigma z^q = 1$. Поэтому каждое линейное отображение $l : \Delta(k) \to R^p$, т. е. такое отображение, что $l(Z) = \sum_{q=1}^{k} z^q l(\Delta_q)$ однозначно определяется набором образов вершин $(l(\Delta_1), \ldots, l(\Delta_k))$. Следовательно, в новой конкретизации множества Z модели 10.3.1 множество L можно отождествить с множеством всех линейных отображений из $\Delta(k)$ в R^p.

В новой интерпретации множеств S_O и L функционал (10.17) можно записать в виде

$$F(s,\ l)=\sum_{q=1}^{k}\sum_{X\in O}\Psi_q(s(X))\,\|X-Y(q)\|^2, \qquad (10.19)$$

где $\Psi_q(s(X))=1$, если $s(X)=\Delta_q$, и 0 — в противном случае, а $Y(q)=l(\Delta_q)$. Тогда согласно общей схеме оптимизационного подхода к построению операторов K и D (см. формулы (10.6) и (10.7)), имеем:

$s_{m+1}=K(s_m,\ l_m)$ — отображение из O в $\Delta(k)$, такое, что

$$s_{m+1}(X)=\arg\min_{s\in S}\sum_{q=1}^{k}\Psi_q(s(X))\,\|X-Y(q)\|^2, \qquad (10.20)$$

т. е. минимум берется по всем возможным значениям классификаций s для данного X. Аргумент, в котором функция из (10.20) достигает минимума, может быть не единственным, поэтому необходимо фиксировать еще способ выбора требуемого аргумента. В формуле для классификатора из алгоритма k-средних (см. п. 7.2.1) заложен в явном виде способ выбора, наиболее употребительный на практике. Далее имеем

$l_{m+1}=D(l_m,\ s_{m+1})$ — линейное отображение из $\Delta(k)$ в R^p, такое, что

$$l_{m+1}(\Delta_q)=\arg\min_{Y\in R^p}\sum_{X\in O}\Psi_q(s_{m+1}(X))\,\|X-Y\|^2,$$

т. е.

$$l_{m+1}(\Delta_q)=\frac{\sum\limits_{X\in O}\Psi_q(s_{m+1}(X))\,X}{\sum\limits_{X\in O}\Psi_q(s_{m+1})(X))}. \qquad (10.21)$$

Итак, требуемая модификация параметров модели 10.3.1 завершена. Она немедленно приводит к модели алгоритма нечеткой классификации, который назвали выше алгоритмом $\Delta(k)$-средних.

Для $Z=\Delta(k)$ возьмем в качестве S_O множество всех отображений из O в $\Delta(k)$. Таким образом, классификация $s\in S_O$ ставит в соответствие объекту X вектор $(z^1,\ \ldots,\ z^k)$, в котором координата z^k имеет смысл степени принадлежности объекта X к q-му классу. Множество L возьмем таким же, как в модели 10.3.1. Интерпретирующий функционал возьмем вида (10.19), но теперь будем считать, что функции $\Psi_q(\cdot)$, $q=1,\ \ldots,\ k$, являются параметрами модели. Операторы K и D зададим согласно формулам (10.20) и (10.21).

10.3.6. Модель алгоритма нечеткой классификации Бездека. Для того чтобы спуститься от модели 10.3.5 к модели 10.3.6, достаточно конкретизировать вид весовых функций $\Psi_q(\cdot)$, рассматриваемых как функции на симплексе $\Delta(k)$. В алгоритме Бездека (см. п. 7.5.2)

$$\Psi_q(Z) = \Psi_q(z^1, \ldots, z^k) = (z^q)^\alpha.$$

В этом случае можно дать явное решение оптимизационной задачи (10.20): найти

$$\operatorname*{argmin}_Z \sum_{q=1}^{k} (z^q)^\alpha \| X - Y(q) \|^2, Z = (z^{(1)}, \ldots, z^k)$$

при условии, что $z^q \geqslant 0$ и $\sum\limits_{q=1}^{k} z^q = 1$. Это решение использовано в описанном (см. п. 7.5.2) алгоритме Бездека.

10.3.7. Модель алгоритма $(\Delta(k) - r)$-средних. Модель 10.3.5 описывает ядерный алгоритм нечеткой классификации, поэтому точно так же, как был осуществлен подъем алгоритма k-средних до алгоритма $(k - r)$-средних, можно от алгоритма $\Delta(k)$-средних перейти вверх по нашей иерархии до алгоритма $(\Delta(k) - r)$-средних. Таким образом получается модель алгоритма, отстоящая от модели 10.3.1 уже на два уровня.

10.3.8. Модель алгоритма $(\Delta(1) - r)$-средних для весовых функций Бездека. Спускаясь вниз по уровням иерархии от модели 10.3.7, получаем:

$Z_O = \Delta(2) = \{Z = z^1\Delta_1 + z^2\Delta_2,\ z^q \geqslant 0,\ z^1 + z^2 = 1\}$;

$S_O = \{s : O \to \Delta(2)\}$; $Y_O = \{ty + (1 - t)Y(*) : 0 \leqslant t \leqslant 1, Y \in R^p\}$ и $Y(*)$ — изолированная точка, т. е. Y_O — конус над R^p с вершиной $Y(*)$. Далее, L — часть множества линейных отображений $\Delta(2)$ в Y, состоящая из всех отображений, при которых вершина Δ_2 переходит в точку $Y(*)$. Классификатор K и дескриптор D задаются формулами:

$$_{m+1}(X) = (s_{m+1,1}(X),\ s_{m+1,2}(X)),\ \text{где}$$

$$s_{m+1,1}(X) = \left[1 + \left(\frac{\| X - Y_m \|}{r}\right)^{\frac{2}{\alpha - 1}}\right]^{-1},\ s_{m+1,2} = 1 - s_{m+1,1},$$

$$l_{m+1}(\Delta_1) = \frac{\sum\limits_{X \in O} (s_{m+1,1}(X))^\alpha X}{\sum\limits_{X \in O} s_{m+1,1}(X)^\alpha}.$$

В заключение отметим, что при $\alpha \to 1$ от модели 10.3.8 переходим (спускаемся по иерархии) к модели 10.3.3, а, полагая

$$\gamma(t) = \left[1 + \left(\frac{t}{r}\right)^{\frac{2}{\alpha - 1}}\right]^{-\alpha},$$

от модели 10.3.4 — к модели 10.3.8.

10.4. Исследование сходимости алгоритмов АК

Определение 10.7. Модель ААК называется *корректной*, если

1) для данных $s \in S_0$, $l \in L$ и $p \in P$ оператор K меняет состояние только подмножества $p \in O$, причем это изменение полностью определяется описанием l;

2) существует интерпретирующий функционал F, ограниченный снизу на данном движении алгоритма, такой, что для любого $\varepsilon > 0$ и $m \geqslant m_0$ существует $\delta > 0$, что, как только $F(s_m, l_m) - F(s_{m+1}, l_{m+1}) < \delta$, то $\rho(l_m, l_{m+1}) < \varepsilon$ (см. определение 10.5).

Далее, говоря об интерпретирующем функционале F корректной модели, всегда будем подразумевать, что он удовлетворяет условию 2 определения 10.7.

Лемма 10.1. Пусть F — интерпретирующий функционал корректной модели ААК. Тогда из равенства $F(s_m, l_m) = F(s_{m+1}, l_{m+1})$ следует, что $l_m = l_{m+1}$ для всех $m \geqslant m_0$.

Оказывается, что для большого класса ААК верно и обратное утверждение.

Лемма 10.2. Если множество значений интерпретирующего функционала F конечно, то в модели ААК условие 2 определения 10.7 выполняется тогда и только тогда, когда из $F(s_m, l_m) = F(s_{m+1}, l_{m+1})$ для $m \geqslant m_0$ следует, что $l_m = l_{m+1}$.

Теорема 10.1. Если множество значений интерпретирующего функционала F корректной модели ААК конечно, то в движении этого алгоритма последовательность описаний $l_1, \ldots, l_m$ стабилизируется, т. е. существует такое m_0, что $l_m = l_{m+1}$ для всех $m \geqslant m_0$.

Доказательство предыдущих результатов использует только свойство 2 определения 10.7. Следующий результат показывает роль свойства 1.

Теорема 10.2. Если в движении алгоритма, описываемого корректной моделью, последовательность они-

саний $l_1, \ldots, l_m$ стабилизируется начиная с номера m_0, то соответствующая последовательность состояний $s_1, \ldots, s_m$ задает стабилизирующуюся классификацию выборки O, т. е. как только $X \in p_n \subset O$ для $m \geqslant m_0$, то $s_{m+1}(X) = s_{m+1+k}(X)$ для всех $k \geqslant 1$.

В случае когда оператор D не зависит от первого аргумента, т. е. задается отображением $S_O \to L$ (это имеет место во многих алгоритмах, например во всех алгоритмах МДС (см. § 7.4)), то из конечности множества S_O и теорем 10.1 и 10.2 следует стабилизируемость движений алгоритмов, описываемых такими корректными моделями. Следующий результат лежит в основе исследования сходимости алгоритмов последовательного типа, описываемых корректной моделью для бесконечных выборок (ср. [139]).

Лемма 10.3. Пусть $(s_1, l_1), \ldots, (s_m, l_m)$ — движение алгоритма, описываемого некоторой корректной моделью. Тогда для любого $\varepsilon > 0$ и натурального N существует номер m_1, такой, что $\rho(l_m, l_{m+t}) < \varepsilon$ для всех $t \leqslant N$ и $m \geqslant m_1$.

Определение 10.8. Модель ААК **называется** *усиленно корректной*, если:

1) оператор K удовлетворяет условию 1 **определения** 10.7;

2) существует интерпретирующий на данном движении алгоритма ограниченный снизу функционал F и строго монотонно возрастающая функция φ, причем $\varphi(0) = 0$, такие, что как только $m \geqslant m_0$, то

$$F(s_m, l_m) \geqslant F(s_{m+1}, l_{m+1}) + \varphi(\rho(l_m, l_{m+1})). \tag{10.22}$$

Нетрудно проверить, что усиленно корректная модель является корректной.

Доказательство усиленной корректности моделей многих ядерных алгоритмов, использующих в качестве меры близости $F_q(O(q), Y)$ между классом $O(q)$ и ядром Y статистический разброс класса $O(q)$ относительно ядра Y, вытекает из классического тождества Гюйгенса:

$$\sum_{X \in O} \gamma(X) \|X - Y\|^2 = \sum_{X \in O} \gamma(X) \|X - \overline{X}\|^2 + \\ + \sum_{X \in O} \gamma(X) |\overline{X} - Y|^2, \tag{10.23}$$

где

$$\overline{X} = \frac{\sum_{X \in O} \gamma(X) X}{\sum_{X \in O} \gamma(X)}.$$

Пример 10.10. Рассмотрим интерпретирующий функционал

$$F(s, l) = \sum_{q=1}^{k} \sum_{X \in O(q)} \| X - Y(q) \|^2$$

в модели алгоритма k-средних. Тогда из (10.23) получаем

$$F(s_m, l_m) \geqslant F(s_{m+1}, l_m) = F(s_{m+1}, l_{m+1}) +$$

$$+ \sum_{q=1}^{k} n_m(q) \| Y_m(q) - Y_{m+1}(q) \|^2,$$

$n_m(q) = | O_m(q) |$, $s_m = (O_m(1), \ldots, O_m(k))$, $l_m =$
$= (Y_m(1), \ldots, Y_m(k))$.

Так как $\rho(l_m, l_{m+1}) = \sum_{q=1}^{k} \| Y_m(q) - Y_{m+1}(q) \|^2$ и $n_m(q) \geqslant 1$ для всех q, то получаем, что модель алгоритма k-средних является усиленно корректной.

Аналогично проверяется, что и модель алгоритма $(k—r)$-средних является корректной для любого r.

Применяя теперь теорему 10.1, получаем, что движение алгоритма $(k - r)$-средних стабилизируется.

Пример 10.11. Пусть $\gamma : R^1 \to [0, 1]$ — невозрастающая функция, являющаяся параметром модели алгоритма выделения размытого кластера (см. п. 10.3.4). Для каждого натурального числа N введем новую функцию $\gamma_N(t) = \frac{[N\gamma(t)]}{N}$, где $[\cdot]$ — символ операции взятия целой части числа. Из свойств операции $[\cdot]$ следует, что $0 \leqslant \gamma(t) - \gamma_N(t) \leqslant \frac{1}{N}$ для всех N.

Теорема 10.3. Допустим, что $\gamma(t) = \lim\limits_{t' \to t-0} \gamma(t')$ для всех $t \in [0, \infty)$. Тогда движение алгоритма выделения размытого кластера с параметром $\gamma_N(t)$ стабилизируется для всех N [42].

В заключение приведем результат, который служит основой для построения новых усиленно корректных моделей ААК.

Лемма 10.4. Пусть заданы следующие компоненты структуры модели ААК: $(S_O, L, P; K, G)$ и некоторый ограниченный снизу функционал $F : S_O \times L \to R^1$, такой, что $F(s, l) \geqslant F(K(s, l, p), l)$. Тогда для любой строго монотонной функции φ существует оператор $D : S_O \times L \to L$, вместе с которым данные компоненты образуют усиленно корректную модель ААК.

Требуемый оператор $D(s, l)$ действует по формуле:

$$D(s, l) = \arg\min_{l' \in L} (F(s, l') + \varphi(\rho(l', l))).$$

ВЫВОДЫ

1. Многообразие алгоритмов автоматической классификации (ААК) представляется в виде некоторой иерархической структуры. На самом верхнем уровне находится математическая модель ААК, компоненты которой образуют средства, облегчающие переход от содержательной постановки задачи классификации к ее математической формализации. На самом нижнем уровне располагаются конкретные алгоритмы АК. Переход с высших уровней на низшие происходит за счет конкретизаций, наполняющих компоненты структуры ААК информацией о характере данных, конечной цели классификации, априорных гипотезах и сведениях о свойствах исследуемых объектов, результатах предварительной обработки и т. п.
2. Описанная теория имеет следующие три аспекта.
А. Сопоставление различных известных алгоритмов и разработка методов целенаправленного конструирования новых алгоритмов, основанных на переходах снизу вверх и сверху вниз по уровням иерархической структуры в многообразии алгоритмов.
Б. Получение параметрических семейств алгоритмов с целью реализации их в виде комплексов программ, предназначенных для автоматизации процедур проверки сложных иерархических гипотез.
В. Исследование модели АК как математической структуры с целью получения условий сходимости алгоритмов и описания класса функционалов, оптимизируемых ими.

Глава 11. ВЫБОР МЕТРИКИ И СОКРАЩЕНИЕ РАЗМЕРНОСТЕЙ В ЗАДАЧАХ КЛАСТЕР-АНАЛИЗА

Проблема выбора метрики и тесно связанная с ней проблема сокращения размерности задачи кластер-анализа возникает, когда исходная информация задана в виде матрицы данных **X**. Выбор метрики, т. е. функции для вычисления расстояния между объектами, является одним из основных управляющих факторов, влияющих на результаты кластер-анализа.

В данной главе рассмотрим несколько подходов, позволяющих в некоторых случаях удовлетворительно решать обе проблемы — выбора метрики и сокращения размерности в тех случаях, когда у исследователя отсутствует априорная информация, позволяющая сделать выбор метрики более обоснованно.

Что касается выделения переменных, то для решения этой задачи в настоящее время не имеется эффективных вычислительных алгоритмов. Частично эта задача решается с помощью процедур адаптивной настройки, менее информативным переменным скорее всего будет присвоен и меньший вес.

11.1. Целенаправленное проецирование данных в пространство небольшой размерности с сохранением кластерной структуры

Этот подход пригоден, когда все переменные измерены в количественной шкале. Будем искать последовательность из q ($q < p$) линейных комбинаций исходных переменных вида $(U_i'X)$, таких, что векторы $U_1, \ldots, U_q$ попарно $\mathbf{S}$-ортогональны и являются решениями оптимизационной задачи

$$U = \arg\max_{\tilde{U}} Q(\tilde{U}, X) \qquad (11.1)$$

при условии $(U_i'\mathbf{S}U_j) = 0$ $(j = 1, \ldots, i-1)$; $\mathbf{S}$ — матрица ковариаций или ее оценка.

В качестве функционала $Q_\beta(U, X)$ используется величина (см. гл. 19)

$$Q_\beta(U, X) = s^\beta \mathrm{E}_f f^\beta(z), \quad \beta > 0,$$

где $f(\ )$ и $s^2(U)$ — соответственно оценки плотности и дисперсии для одномерной случайной величины $z = U'X$, оцененной по совокупности одномерных проекций $z_l = (U' X_l)$ $(l = 1, \ldots, n)$.

Смысл использования критерия (11.1) состоит в том, что чем больше его величина, тем более неоднородным можно считать распределение одномерной проекции $z = (U' X)$, например, в рамках модели смеси нормальных распределений.

Перейдем сначала к махаланобисовой метрике, т. е. сделаем преобразование $Y = \mathbf{S}^{-1/2}X$. Пусть из условия максимума (11.1) определены линейные комбинации $U_1, \ldots, U_q$. Теперь они будут ортогональны, так как в новом базисе $\mathbf{S} = \mathbf{I}_p$. И пусть $Q_1, \ldots, Q_q$ — соответствующие им значения

функционала $Q_\beta(U, X)$. Вместо исходного p-мерного признакового пространства будем далее использовать q-мерное пространство новых переменных $z^{(i)} = (U_i' X)$, предварительно нормированных так, чтобы $s^2(z^{(i)}) = 1$ (при этом величина Q_i не меняется). Расстояние между объектами будем вычислять следующим образом:

$$d_{ij}^2 = \sum_{k=1}^{q} v_k \left(z_i^{(k)} - z_j^{(k)}\right)^2,$$

где $z_i^k = (U_k' X_i)$, $v_k = \varphi(Q_k)$,

$\varphi(\cdot)$ — некоторая монотонно возрастающая функция.

11.2. Метрики для задач кластер-анализа с неколичественными переменными

Некоторые из метрик для измерения расстояний между объектами, когда переменные являются неколичественными, приведены в гл. 5. Из них наиболее простой является хэммингова метрика, которую можно определить как

$$d_h^2(X_k, X_l) = \frac{\text{число переменных, у которых объекты } X_k, X_l \text{ попали в разные категории}}{\text{(общее число переменных)}} \qquad (11.2)$$

Расстояние Хэмминга можно рассматривать как квадрат евклидова расстояния в пространстве бинарных переменных, соответствующих категориям исходных переменных (далее, для краткости, просто в пространстве категорий), т. е.

$$d_h^2(X_k, X_l) = \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{l_i} \left(y_{kj}^i - y_{lj}^i\right)^2, \qquad (11.2')$$

где i — номер исходной переменной; j — номер категории; l_i — число категорий i-й переменной.

Иногда хэммингową метрикой называют величины (11.2) и (11.2'), у которых отсутствует деление на p.

Так как величины y_{kj}^i могут принимать лишь значения 1 (для k-го объекта реализовалась j-я категория i-й переменной) или 0 (в противном случае), то выражения (11.2) и (11.2') совпадают.

Теперь, по аналогии с евклидовой метрикой, можно подчеркнуть важность переменных или отдельных их категорий в формировании различий между объектами, вводя веса либо для переменных, либо даже для отдельных категорий (т. е. бинарных переменных y_j^i).

Один из подходов к присваиванию весов ω_j^i категориям состоит в переходе к χ^2-метрике, возникающей в множественном анализе соответствий. Веса для категорий в этой метрике возникают при решении оптимизационной задачи, имеющей ясную статистическую интерпретацию (см. п. 17.2.5), а не внесены извне. Поэтому можно полагать, что χ^2-метрика определяет некоторую «естественную» меру измерения отношений между объектами и, следовательно, ее целесообразно использовать при проведении кластер-анализа в качестве одного из основных претендентов.

Другой способ введения весов, основанный на эвристических соображениях, предложен в работе [174].

Пусть для i-й переменной в категорию j попало n_j объектов. Тогда для двух случайно выбранных объектов определим вероятности следующих событий:

у обоих объектов одна и та же j-я категория i-й переменной

$$P_{jj}^i = P\{y_{kj}^i y_{lj}^i = 1\} = (n_j^i/n)^2;$$

у k-го объекта реализовалась категория j, а у l-го — категория r

$$P_{jr}^i = P\{y_{kj}^i y_{lr}^i = 1\} = 2n_j^i n_r^i/n^2.$$

Будем вводить веса категорий исходя из следующего соображения. Пусть для i-го признака для k-го объекта $y_{kj}^i = 1$ (реализовалась j-я категория), а для l-го объекта — $y_{lr}^i = 1$. Чем меньше вероятность P_{jr}^i такого события при случайном выборе объектов, тем более близкими их будем считать [1]. Чтобы получить теперь расстояние для объектов, можно воспользоваться следующим подходом. Определим меру близости между объектами в виде

$$\Delta^2 = \sum_{i=1}^{p} \sum_{j,\, r} \omega_{jr}^i y_{kj}^i y_{lr}^i. \tag{11.3}$$

Вклад i-й переменной в Δ^2

$$\Delta_i^2 = \sum_{j,\, r}^{l_i} \omega_{jr}^i y_{kj}^i y_{lr}^i, \tag{11.4}$$

где

$$\omega_{jr}^i = (2/P_{jr}^i)/[l_i(l_i + 1)].$$

[1] Если оба объекта принадлежат к одной редкой группе (категории), то это может оказаться более важным, чем сходство или различие по другим переменным.

Так как только одно из произведений $y_{kj}^i y_{lr}^i$ отлично от нуля, а все остальные равны нулю, то реально вклад Δ_{ij}^2 равен одному из весов ω_{jr}^i. Это взвешивание как раз и увеличивает сходство согласно вышеизложенному принципу — чем меньше вероятность реализованной комбинации категорий переменной для наблюдаемых двух объектов, тем больше сходство между этими объектами.

Выражение (11.3) есть не что иное, как скалярное произведение вида

$$\Delta^2 = Y_k' \mathbf{W} Y_l, \tag{11.5}$$

где матрица $\mathbf{W}$ — блочно-диагональная матрица весов ω_{jr}^i.

Евклидово расстояние из Δ^2 можно теперь получить, используя обычную формулу

$$d^2(X_k, X_l) = \|Y_k\|_{\mathbf{W}}^2 - 2\Delta^2 + \|Y_l\|_{\mathbf{W}}^2, \tag{11.6}$$

где

$$\|Y_k\|_{\mathbf{W}}^2 = Y_k' \mathbf{W} Y_k = \sum_{i=1}^{p} \sum_{j=1}^{l_i} \omega_{jj}^i y_{kj}^i.$$

Для введения метрики в пространстве неколичественных переменных можно использовать подход, основанный на оцифровке, т. е. присвоении меток неколичественным переменным, например по критерию (17.31) (см. § 17.3).

11.3. Алгоритмы классификации с адаптивной метрикой

Один из способов получения метрики, подходящей для классификации, состоит в ее уточнении («настройке») в процессе работы самой процедуры классификации.

11.3.1. Адаптивная махаланобисова метрика. Квадрат расстояния между точками x_i и x_j задается в этом случае (см. гл. 5) как

$$d^2(X_l, X_m) = d_{lj}^2 = (X_l - X_m)' \mathbf{V} (X_l - X_m), \tag{11.7}$$

где $\mathbf{V}$ — некоторая положительно-определенная симметричная матрица. Для определенности положим, что определитель $\mathbf{V}$ равен 1, $\det(\mathbf{V}) = 1$.

Докажем следующую лемму.

Лемма 11.1. Пусть расстояние между точками задано в виде (11.7); точки распадаются на k непересекающихся кластеров $G_1, \ldots, G_k$; $\mathbf{S}$ — матрица полного рассеивания (ковариационная матрица для X).

Пусть теперь W — величина внутриклассового рассеивания, вычисленная на основе расстояния (11.7),

$$W(\mathbf{V}) = \sum_{i=1}^{k} \sum_{X_l,\ X_m \in G_i} d^2(X_l,\ X_m). \tag{11.8}$$

Минимальное значение $W(\mathbf{V})$ достигается на множестве положительно-определенных матриц $\mathbf{V}$ с единичным определителем тогда, когда матрица $\mathbf{V} = \alpha \mathbf{W}^{-1}$, где $\mathbf{W}$ — матрица внутриклассового разброса

$$\mathbf{W} = \frac{1}{n} \sum_{i=1}^{k} \sum_{X_l,\ X_m \in G_i} (X_l - X_m)(X_l - X_m)', \tag{11.9}$$

а множитель α выбран таким, чтобы $|\mathbf{V}| = \alpha |\mathbf{W}^{-1}| = 1$, т. е. $\alpha = |\mathbf{W}|^{1/p}$.

Доказательство. Величину внутриклассового разброса можно представить в виде

$$W(\mathbf{V}) = \sum_{i=1}^{k} \sum_{X_l,\ X_m \in G_i} d_{lm}^2 = \sum_{i=1}^{k} \sum_{X_l,\ X_m \in G_i} (X_l - X_m)' \mathbf{V} \times$$

$$\times (X_l - X_m) = \sum_{i=1}^{k} \sum_{X_l,\ X_m \in G_i} \mathrm{Sp}\,(\mathbf{V}(X_l - X_m)(X_l - X_m)') =$$

$$= n\,\mathrm{Sp}\,(\mathbf{VW}). \tag{11.10}$$

Дальше доказательство аналогично доказательству теоремы в приложении 1 [106, гл. 12].

Рассмотрим теперь следующий двухфазный алгоритм классификации

Фаза 1. При фиксированной метрике $\mathbf{V}^{(t)}$ (t — номер шага итерации) проводим разделение выборки $\mathbf{X}$ с помощью алгоритма k-средних (Мак-Кина). Число классов k задается пользователем при начале работы алгоритма и дальше не меняется.

Фаза 2. По полученной классификации вычисляем матрицу внутриклассового разброса $\mathbf{W}^{(t+1)}$ согласно формуле (11.9). Вводим новую метрику с матрицей

$$\mathbf{V}^{(t+1)} = (\mathbf{W}^{(t+1)})^{-1}$$

(множитель α_{t+1} вычисляется попутно в процесс обращения матрицы $\mathbf{W}$, но, вообще говоря, его использование необязательно) и переходим к фазе 1.

Остановка алгоритма производится либо когда матрица $\mathbf{V}^{(t)}$ перестанет изменяться, либо когда относительное умень-

шение критерия w_t станет меньше пороговой величины, т. е. либо

$$\| \mathbf{V}^{(t+1)} - \mathbf{V}^{(t)} \| < \varepsilon_{\mathbf{V}},$$

либо $(W^{(t)} - W^{(t+1)}) < \varepsilon_{\mathbf{W}}$.

Сходимость алгоритма (в вычислительном отношении) следует из того, что на каждой фазе работы алгоритма значение критерия W убывает.

11.3.2. Состоятельность алгоритма с адаптивной махаланобисовой метрикой. Рассмотрим следующую вероятностную модель — смесь эллипсоидально-симметричных распределений (см. гл. 19) с ограниченными непересекающимися носителями и одинаковой ковариационной матрицей $\mathbf{W}$.

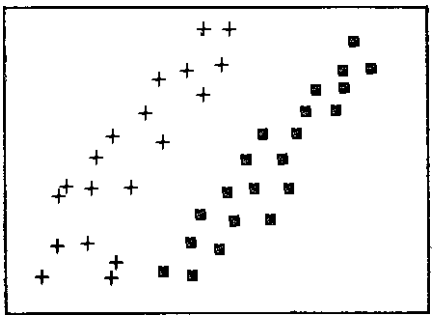


Рис. 11.1. Классификация данных табл. 11.1

Тогда, если $\mathbf{X}$ — выборка из такого распределения, то верна следующая лемма.

Лемма 11.2. Алгоритм с адаптивной махаланобисовой метрикой есть ЕМ-алгоритм (см. гл. 6) решения уравнения правдоподобия для оценки параметров смеси: набора весов a_i ($i = \overline{1, k}$), средних векторов компонент M_i и матрицы ковариаций $\mathbf{W}$. Соответствующими оценками будут величины n_i/n, $\overline{X}_i$, $\mathbf{W}$ (11.9), оцененные на последнем шаге работы алгоритма. Поскольку ЕМ-алгоритм дает оценку максимального правдоподобия, то состоятельность получаемых оценок (при $n \to \infty$) следует из общей теории.

Пример 11.1. Рассмотрим набор двумерных данных из п. 12.5.2 (табл. 11.1).

Результаты применения алгоритма из п. 11.3.1 представлены на диаграмме рассеивания (рис. 11.1) («крестик» — точки, отнесенные в 1-й кластер, «квадрат» — точки, отнесенные во 2-й кластер). Состав кластеров: первые 20 точек из табл. 11.1 — первый кластер, точки с 21 по 40 — второй. Как можно судить по рисунку, обычная евклидова метрика здесь не дала бы успешной классификации.

Таблица 11.1

Номер объекта	$x^{(1)}$	$x^{(2)}$	Номер объекта	$x^{(1)}$	$x^{(2)}$
1	40,89	35,00	21	29,93	22,05
2	39,04	36,51	22	25,62	22,05
3	38,49	34,79	23	25,68	21,53
4	38,08	33,49	24	25,00	25,00
5	37,40	31,85	25	20,21	21,58
6	35,89	31,95	26	23,01	24,45
7	36,71	33,49	27	21,03	24,45
8	35,66	30,55	28	25,14	26,99
9	34,11	30,55	29	23,29	26,99
10	35,27	29,11	30	23,68	30,96
11	33,42	29,11	31	24,04	29,73
12	33,49	27,74	32	26,92	29,73
13	31,92	27,74	33	27 33	33,70
14	35 55	27,74	34	29,38	33,70
15	34,45	25,62	35	28,42	31,99
16	32,33	25,62	36	30,21	36,44
17	30,41	25,62	37	28,36	36,44
18	29,86	23 97	38	25,96	33,08
19	31,78	23,97	39	21,71	27,47
20	28,01	22,05	40	23,42	28,90

Исходная матрица

$$\mathbf{V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Матрица $\mathbf{V}$ после четырех итераций (конец работы алгоритма) практически совпала с результирующей матрицей, полученной в [106].

$$\mathbf{V} = \begin{bmatrix} 2,66 & -1,52 \\ -1,52 & 1,19 \end{bmatrix}.$$

11.3.3. Адаптивная взвешенная евклидова метрика. Предположим, что матрица $\mathbf{V}$ в (11.7) диагональна: $\mathbf{V} = \operatorname{diag}(v_1^2, \ldots, v_p^2)$. Тогда (11.7) соответствует взвешенной евклидовой метрике. Действуя так же, как в лемме 11.1, можно показать, что веса v_r^2, минимизирующие внутриклассовый разброс $W(\mathbf{V})$, определяются соотношениямн

$$v_r^2 = \alpha w^{rr}, \tag{11.11}$$

где w^{rr} — r-й днагональный элемент матрицы $\mathbf{W}^{-1}$, а α — нормирующий множитель — выбирается так, чтобы $\prod_{r=1}^{p} v_r^2 = 1$ (выполнение условия $|\mathbf{V}| = 1$).

Состоятельность алгоритма, описанного в п. 11.3.1, будет теперь иметь место только в случае модели смеси распределений с диагональной внутрикомпонентной матрицей рассеивания **W** (см. лемму 11.2). Поэтому на практике можно заменить (11.11) более простым выражением

$$v^2 = \alpha w_{rr}^{-1}. \tag{11.12}$$

Состоятельность для смеси с диагональной внутрикомпонентной матрицей имеет место и в этом случае, а точность оценок даже будет лучше.

В случае недиагональной матрицы **W** оба способа вычисления весов приведут к смещенным оценкам параметров смеси.

Вычислительная сходимость алгоритма с адаптивной диагональной матрицей следует из тех же соображений, что и в п. 11.3.1.

В отличие от махаланобисовой метрики, результаты классификации в данном случае зависят от исходной метрики.

11.3.4. Адаптивная взвешенная метрика типа «сити-блок». Расстояние в этой метрике определяется как

$$d(X_l, X_m) = \sum_{j=1}^{p} v_j \left| x_l^{(j)} - x_m^{(j)} \right|.$$

Потребуем, чтобы $\prod_{j=1}^{p} v_j = 1$.

Алгоритм снова состоит из двух фаз, как и в п. 11.3.1, но имеются следующие отличия:

1) центр i-го класса $\overline{X}_i$ определяется как вектор, компоненты которого суть *медианные* значения признаков в i-м классе;

2) внутриклассовый разброс находится по формуле

$$W(V) = \sum_{i=1}^{k} \sum_{X \in G_i} d(X, \overline{X}_i). \tag{11.13}$$

На фазе 2 вектор весов V, минимизирующий $W(V)$, определяется (см. [106, п. 12.4.2.2]) из следующего выражения

$$v_j = \frac{\prod_{h=1}^{p} \left(\sum_{i=1}^{k} \sum_{X \in G_i} \left| x^{(h)} - \overline{x}_i^{(h)} \right| \right)^{1/p}}{\sum_{i=1}^{k} \sum_{X \in G_i} \left| x^{(j)} - \overline{x}_i^{(j)} \right|}. \tag{11.14}$$

Заметим, что в [106] приведены выражения для v_j, т. е. метрика считается разной в разных классах. Здесь приведен вариант весов, полученный в предположении одинаковости весов во всех классах.

11.4. Оценка метрики с помощью частично обучающих выборок

Понятие частично обучающей выборки (ЧОВ) введено в работе [9, гл. 1]. ЧОВ определяется как множество пар объектов, таких, что относительно двух объектов, составляющих некоторую пару, известно, что они принадлежат одному и тому же классу. Более детальная информация, вообще говоря, отсутствует. Например, неизвестно, принадлежат ли некоторые пары, составленные из непересекающихся пар объектов, одному и тому же классу или нет. Таким образом, фактически исследователь на примерах определяет, какие объекты считать близкими, если исходить из неформализованных содержательных представлений.

Пусть дальше $n_{\text{чов}}$ — число пар в ЧОВ, а n_0 — число независимых объектов, входящих в множество пар из ЧОВ.

Рассмотрим теперь следующий способ оценки метрики, основываясь на ЧОВ. Предположим, что неизвестная нам метрика является взвешенной евклидовой $d^2_{ij} = \sum_{k=1}^{p} v_k^2 (x_i^{(k)} - x_j^{(k)})^2$, причем все веса $v_k^2 > 0$ (ненулевые).

Без ограничения близости можно считать, что выполняется условие

$$\prod_{i=1}^{p} v_i^2 = 1. \tag{11.15}$$

Выполнения этого условия можно добиться, умножая все веса v_i на одно и то же положительное число α, т. е. одновременно и одинаково изменяя масштаб по всем переменным. Это, естественно, не влияет на результаты применения кластер-процедур.

Суммируя расстояния между всеми парами, из ЧОВ получаем

$$W(\mathbf{V}) = \operatorname{Sp} \mathbf{WV}, \tag{11.16}$$

где $\mathbf{V} = \operatorname{diag}(v_1^2, \ldots, v_p^2)$;

$$\mathbf{W} = \sum_{i=1}^{n_{\text{чов}}} (X_{1i} - X_{2i})(X_{1i} - X_{2i})' \ (X_{1i}, X_{2i} - i\text{-я пара из ЧОВ}).$$

Так как слагаемые в (11.16) суть расстояния между парами точек из одного и того же класса, т. е. близкими между собой точками, нужно стремиться получить $\mathbf{V}$, такую, чтобы значение (11.16) было как можно меньше (при выполнении условия (11.15)). Итак, веса $\mathbf{V}$ — это решение минимизационной задачи

$$\operatorname{Sp} \mathbf{WV} \Rightarrow \min_{\mathbf{V}} \tag{11.17}$$

при условии $\prod_{k=1}^{p} v_k^2 = 1$.

Решением задачи (11.17) будут следующие значения весов (см. п. 11.3.3):

$$v_i^2 = \alpha w_{ii}^{-1}, \tag{11.18}$$

где значение параметра $\alpha > 0$ выбирается так, чтобы удовлетворялось условие (11.15). Впрочем, выбор α несуществен, поскольку задача кластер-анализа инвариантна относительно изотропного одновременного изменения масштаба переменных.

Если объем ЧОВ достаточно велик, чтобы матрица $\mathbf{W}$ была невырождена, то можно построить и оценку махалапобисовой метрики, решая задачу (11.17), но уже не считая матрицу $\mathbf{V}$ диагональной. Решением будет матрица $\mathbf{V} = \alpha \mathbf{W}^{-1}$, а метрика будет задаваться выражением

$$d_{ij}^2 = (X_i - X_j)' \mathbf{V} (X_i - X_j).$$

ВЫВОДЫ

1. В случае когда у исследователя отсутствует априорная информация о том, как измерять расстояния между объектами в пространстве переменных, и шкалы, в которых измерены переменные, количественные, полезными могут оказаться предварительное сокращение размерности пространства с помощью методов целенаправленного проецирования (подробнее см. гл. 19) и конструирование метрики в пространстве сокращенной размерности. Этот подход не следует использовать, когда объем выборки невелик ($n < 100$ или $p/n > 0{,}5$).
2. В случае неколичественных переменных можно сконструировать метрики, являющиеся взвешенными вариантами метрики Хэмминга. Среди них особого внимания заслуживает метрика χ^2 (см. 17.4).

3. Целесообразно использование алгоритмов с адаптивной метрикой (§ 11.3).
4. При наличии некоторого типа априорной информации о близостях между объектами частично обучающих выборок оказывается возможным оценить весовые коэффициенты для адекватной взвешенной евклидовой метрики, а при достаточном объеме информации — и матрицу метрики махаланобисова типа. Используя эти оценки как стартовые, можно затем применить для их уточнения алгоритмы с адаптивной метрикой.

Глава 12. СРЕДСТВА ПРЕДСТАВЛЕНИЯ И ИНТЕРПРЕТАЦИИ РЕЗУЛЬТАТОВ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

12.1. Некоторые средства оценки результатов кластер-анализа

12.1.1. Оценка качества классификации с помощью критериев классификации. Предположим, что, используя некоторую процедуру кластер-анализа (классификации), получили разбиение объектов из нескольких групп. Один из важных вопросов, который возникает у исследователя: насколько удачно полученное разбиение. Основным критерием качества и обоснованности полученного разбиения является содержательный анализ результатов, основанный на осмыслении исследователем возможных причинных механизмов осуществления и обособления полученных групп объектов. Чисто статистические критерии оказывают лишь помощь в этом процессе. С одной стороны, они позволяют отбраковывать плохие группировки, но, с другой стороны, группировка, удачная по этим критериям, может и не иметь содержательной ценности.

Известны десятки критериальных величин, используемых в кластер-анализе (см. гл. 5, 7, 10, 11). В работе [273] тридцать из них подвергнуто изучению методом статистического моделирования. В результате эти критерии были упорядочены по степени согласованности их величины с удачностью применения кластерного анализа (использовалось 15 различных процедур) к массивам данных, кластерная структура которых была заранее известна. Две величины, которые рассматриваются дальше, входили в шестерку лучших. Следует отметить, однако, что при проведении моделирования использовалась только евклидова метрика. В част-

ности, возможно, поэтому инвариантные критерии не «проявили» себя в должной мере и не попали в шестерку лучших.

Пусть совокупность объектов разбита на k групп $G_1, \ldots, G_k$.

Рассмотрим здесь следующие две величины, полезные для оценки качества разбиения: величина объясненной доли общего разброса T и точечно-бисериальный коэффициент корреляции R_b. Некоторые другие величины приведены также в § 12.2.

Чтобы определить величину T, введем следующие три характеристики степени рассеивания объектов из $\mathbf{X}$:

общее рассеивание $S = \sum_{i=1}^{n} d^2(X_i, \overline{X})$; (12.1)

межклассовый разброс $B = \sum_{j=1}^{k} n_j d^2(\overline{X}_j, \overline{X})$; (12.2)

внутриклассовый разброс $W = \sum_{j=1}^{k} W_j$, $W_j =$

$$= \sum_{X_i \in G_j} d^2(X_i, \overline{X}), \tag{12.3}$$

где $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ — общий центр тяжести, $\overline{X}_j = \frac{1}{n_j} \sum_{X_i \in j} X_i$ — центр тяжести j-й группы[1]; n_j — число объектов в группе G_j.

Если используется евклидово или взвешенное евклидово расстояние, то имеет место известное равенство

$$S = W + B. \tag{12.4}$$

Рассмотрим величину

$$T = 1 - W/S. \tag{12.5}$$

Чем больше величина T, тем большая доля общего разброса точек «объясняется» межклассовым разбросом и можно считать, с определенным основанием, тем лучше качество разделения. Очевидно, $0 \leqslant T \leqslant 1$.

Точечно-бисериальный коэффициент корреляции R_b определяется следующим образом. Каждой паре объектов X_i и X_j поставим в соответствие две величины — расстояние

[1] Если входной является матрица расстояний, то в качестве центра группы выбирается объект (строка матрицы $\mathbf{D}$) X_l, такой, что если $\overline{X} = X_l$, величина S принимает минимальное значение. Из аналогичных соображений выбираются центры групп.

между ними в выбранной метрике и индекс эквивалентности

$$\delta_{ij} = \begin{cases} 1, \text{ если } X_i \text{ и } X_j \text{ принадлежат одному классу;} \\ 0 \text{ — в противном случае.} \end{cases}$$

Коэффициент R_b подсчитывается как обычный коэффициент корреляции между d_{ij} и бинарной величиной δ_{ij} по всем парам объектов, что дает

$$R_b = (\bar{d}_b - \bar{d}_w)(f_w f_b / n_d^2)^{1/2} / s_d \qquad (12.6)$$

где $\bar{d}_b$ — среднее расстояние между точками из разных кластеров;

$\bar{d}_w$ — среднее расстояние между точками из одного кластера;

f_w — число расстояний между точками, попавшими в одну группу;

f_b — число расстояний между точками из разных кластеров;

n_d — общее число расстояний;

s_d — стандартное отклонение расстояний.

12.1.2. Оценка компактности выделенных групп. Другие полезные для оценки качества разбиения характеристики можно ввести с помощью следующих определений [110].

Кластером называется группа объектов G_i, такая, что выполняется неравенство $\bar{d}_i^2 = W_i/n < S/n$, т. е. средний квадрат внутригруппового расстояния до центра группы меньше среднего квадрата расстояния до общего центра в исходной совокупности. Чем больше среди групп G_j кластеров, тем более успешным можно считать разбиение.

Еще более полезным является понятие «сгущение». Группа объектов G_i называется *сгущением*, если максимальный квадрат расстояния объектов из G_i до центра группы меньше $\bar{d}^2 = S/n$, т. е. $d^2_{i,\max} = \max\limits_{X_j \in G_i} d^2(\bar{X}_i, X_j) < \bar{d}^2$.

В [110] эти понятия введены в случае, когда используются не расстояния между объектами, а некоторые меры близости между ними.

Агломеративные иерархические процедуры классификации устроены так, что группировки, получаемые при разрезании дерева на любом уровне, будут кластерами в смысле, определенном выше. Для других процедур, например типа k-средних, это не гарантируется, поэтому получение кластеров при их применении можно рассматривать как достаточно важное указание на хорошее качество разделения.

12.1.3. Визуальные средства оценки степени разнесенности и компактности выделенных групп объектов.

Полезным средством, позволяющим быстро оценить успешность разделения, компактность классов, наличие в них выбросов и т. д., являются одно-, двумерные отображения множества точек, с указанием их групповой принадлежности, в виде гистограмм и диаграмм рассеивания на некоторые подходящим образом выбранные направления. В качестве таких отображений обычно используют отображения на оси главных компонент и факторные оси (количественные признаки, см. гл.13):

нелинейное отображение (**количественные переменные**, см. гл. 13);

метрическое и неметрическое шкалирование (обрабатывается матрица расстояний или удаленностей, см. гл. 16);

оси, получаемые в анализе соответствий (неколичественные переменные и переменные смешанной природы, см. § 17.2).

В случае количественных, а также оцифрованных (§ 17.3) переменных эффективным будет отображение на канонические дискриминантные направления (подробнее о них см. гл. 19), которые определяются как собственные векторы обобщенной задачи на собственные числа и векторы вида $(\mathbf{S}-\lambda\mathbf{W})=0$, где $\mathbf{S}$ — полная ковариационная матрица; $\mathbf{W}$ — матрица внутригруппового разброса.

Для получения проекций используются векторы, соответствующие наибольшим собственным значениям. Заметим, что имеется не более чем $\min(p, k-1)$ ненулевых собственных чисел. Отсюда следует, что если имеется $k=3$ класса и $p>2$, то отображение на плоскость двух первых канонических направлений содержит полную информацию о различиях между классами, и их образы необходимо не должны иметь пересечения (если в процедурах классификации использовалась какая-то разновидность евклидовой метрики). При $k>3$, вообще говоря, отображение на плоскость, определяемую первыми двумя каноническими направлениями, может содержать пересекающиеся классы, их отсутствие возможно только при определенном расположении центров тяжести классов (на некоторой плоскости, на прямой или на плоской кривой в p-мерном пространстве; см., например, рис. 12.1).

Но тогда можно использовать и больше канонических векторов и исследовать отображения, например, на 1-е и 3-е или 2-е и 3-е направления и т. д. Отображение, определяемое парой канонических направлений, на котором любой указанный класс будет отделен от других, должно существовать

Из сказанного, в частности, можно сделать следующий вывод. Если на плоскости, определяемой первыми двумя ка-

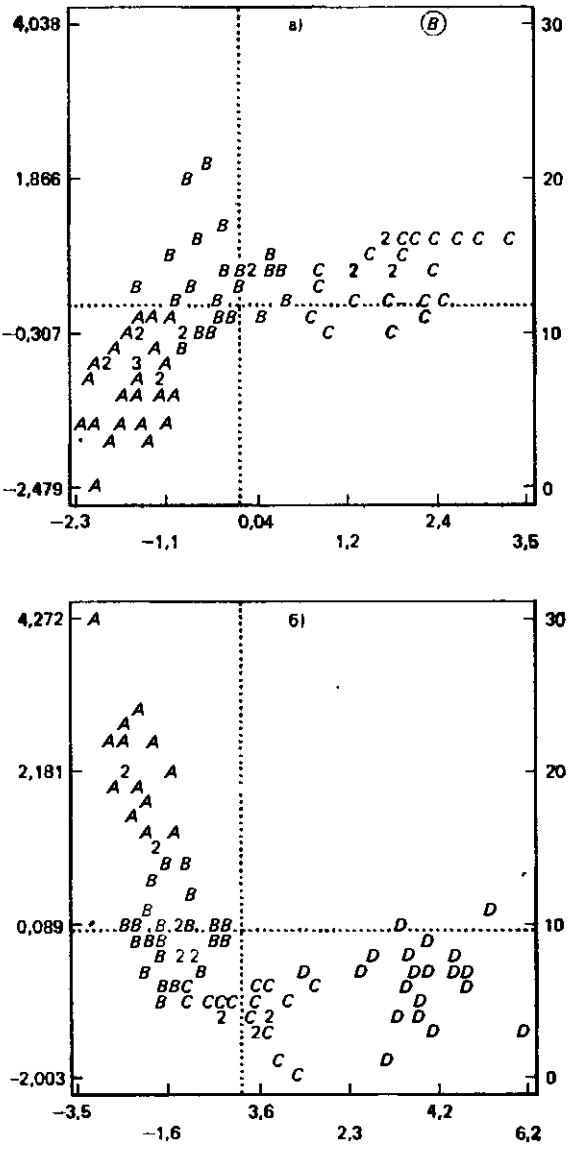


Рис. 12.1. Отображение результатов классификации на плоскость двух первых канонических направлений
а) $k=3$;
б) $k=4$

ноническими направлениями, разделены все группы и $k > 3$, то это означает определенную закономерность в расположении центров классов и, следовательно, уверенность в том, что это деление несет в себе некоторую смысловую нагрузку, возрастает.

Отображения можно использовать для нескольких целей. Во-первых, для получения перечисленной в начале параграфа информации. Во-вторых, для получения информации о структуре, которую образуют сами кластеры, например, об их возможной пространственной упорядоченности, имеющей в то же время содержательный смысл, как это видно из примера 12.1 (рис. 12.1). Такую информацию трудно получить другими способами. В-третьих, для интерпретации. Поскольку в большинстве случаев (за исключением нелинейного отображения и шкалирования) отображения определяются векторами, коэффициенты этих векторов можно использовать для интерпретации таким же способом, как и нагрузки в факторном анализе.

Пример 12.1. Применим процедуру классификации (разделения смесей) к реальным данным[1]. Матрица этих данных содержит значения 31 показателя социально-экономического развития для 85 несоциалистических стран (данные относятся к началу 70-х годов). Из этих переменных нами было использовано 29.

Приведем в сокращенном виде результаты работы программы при разбиении на три класса

1-Й КЛАСС (ГРУППА)
НОМЕРА ОБЪЕКТОВ 59 60 21 51 82 30 74 81 9 ...
КОЛИЧЕСТВО ОБЪЕКТОВ В КЛАССЕ = 31
2-Й КЛАСС (ГРУППА)
НОМЕРА ОБЪЕКТОВ 64 84 11 58 42 37 47 44 39 ...
КОЛИЧЕСТВО ОБЪЕКТОВ В КЛАССЕ = 27
3-Й КЛАСС (ГРУППА)
НОМЕРА ОБЪЕКТОВ 4 78 36 73 52 1 50 20 6 77 ...
КОЛИЧЕСТВО ОБЪЕКТОВ В КЛАССЕ = 27
СУММА РАССТОЯНИЙ ДО ОБЩЕГО ЦЕНТРА $S = 40\,649$
СРЕДНЕЕ РАССТОЯНИЕ ДО ОБЩЕГО ЦЕНТРА $S/N = 4.782$
ДОЛЯ РАЗБРОСА, ОБЪЯСНЕННАЯ КЛАССИФИКАЦИЕЙ, $T = 0.270$.
БИСЕРИАЛЬНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ $R_b = 0.404$

[1] Типология несоциалистических стран — М.: Наука, 1976 — 272 с.

Часть результатов, полезных для анализа удачности разбиения, суммируется в табл. 12.1.

Таблица 12.1

Номер группы	Среднее расстояние до центра $(\overline{d}_i^2)$	Эталонный объект	Максимальное расстояние до центра $(d^2_{i,\ \max})$	Объект, наиболее удаленный от центра
1	2,769	54	4,37	12
2	3,343	57	14,43	64
3	4,369	76	7,975	40

В третьем столбце приведены номера эталонных объектов, наиболее близких к центрам групп, в пятом — номера объектов, на которых достигаются максимальные расстояния.

Согласно определению, приведенному в п. 12.1.2, все три выделенные группы являются кластерами, а первая будет также сгущением Значения критериев Т и В также достаточно велики. Однако визуальный анализ рис. 12.1а (проекции на канонические направления) показывает, что разделение групп 1 и 2 (символы A и B соответственно) нельзя признать выраженным. Скорее можно считать, что существует непрерывный переход от группы A к группе B. На рисунке хорошо выделен один объект из 2-й группы (обведен кружком), на котором реализуется максимальное расстояние. Группа 3 (символ C) хорошо отделена от первых двух групп.

Применим тот же алгоритм к тем же данным, но положим $k=4$, т. е. проводим разделение на 4 группы. Результаты классификации теперь будут такими: $T = 0{,}356$, $R_b = 0.450$. Другие величины приведены в табл. 12.2.

Таблица 12.2

Номер группы	$\overline{d}_i^2$	$d^2_{i,\ \max}$	Объем группы
1	2,363	4,605	17
2	2,524	5,084	29
3	3,336	5,933	19
4	4,254	9,208	20

Снова все выделенные группы — кластеры, а одна из них — сгущение. Значения d_i^2 и $d_{i,\max}^2$ уменьшились. Визуальная картина разделения (см. рис. 12.1б) также указывает на лучшее качество разделения между группами и отсутствие далеко отстоящих объектов. На рисунке ясно прослеживается подковообразная структура, образованная проекциями объектов. Следует отметить, что на рис. 12.1 а и б символы, соответствующие одному и тому же объекту, могут не совпадать. Так, большинство объектов из группы 2 примера 12.1а (символ *B* на рис. 12.1а) перешли в группу 4 (символ *D* на рис. 12.1б). Соотнесение объектов из выделенных групп ($k = 4$) с исходными данными показывает, в частности, что в группу *D* вошли в основном высокоразвитые страны (США, европейские страны, Япония) и, напротив, в группу *A* — развивающиеся страны с низкими показателями социально-экономического развития. Таким образом, расположение кластеров, упорядоченное вдоль указанной подковообразной кривой, соответствует их некоторому содержательному упорядочению, что, по-видимому, повышает доверие к результату классификации.

12.2. Связь между показателями качества прогноза переменных, метрикой и некоторыми критериями качества классификации в кластер-анализе

12.2.1. Случай, когда переменные измерены в количественной шкале. Рассмотрим задачу кластер-анализа (классификации) в формулировке, обобщающей постановку задачи отыскания главных компонент. Будем искать номинальную категоризованную переменную (фактор) *z*, имеющую *k* категорий, такую, чтобы критерий вида

$$K^2 = \sum_{i=1}^{p} v_i \rho^2(x^{(i)}, z) \Rightarrow \max_z, \qquad (12.7)$$

где $\rho^2(x^{(i)}, z)$ — корреляционное отношение (см., например, [7,12]) между $x^{(i)}$ и *z*; v_i — весовой коэффициент, $0 \leqslant v_i \leqslant 1$. Иными словами, нужно получить такую классификацию объектов, которая наилучшим образом, в смысле критерия (12.7), объясняла бы наблюдающийся разброс переменных $x^{(i)}$ ($i = \overline{1, p}$). Вес же v_i определяет степень важности, которую придаем «объяснению» переменной $x^{(i)}$ посредством фактора *z*. Такого рода группировки объектов

в [110] предлагается называть объясняющими (фактор z «объясняет» переменные $x^{(i)}$).

Далее будем полагать, что матрица данных $\mathbf{X}$ центрирована. Классификационную переменную z можно представить в виде булевой матрицы $\mathbf{Z}$ с k столбцами и n строками и такой, что элемент $z_{ij} = 1$, если i-й объект принадлежит j-му классу (j-й категории переменной z) и $z_{ij} = 0$ в противном случае.

Такое представление часто используется, например, в регрессионном и дисперсионном анализе для введения так называемых фиктивных переменных. Столбцы матрицы $\mathbf{Z}$ ортогональны

$$Z'_{\cdot j} Z_{\cdot i} = \begin{cases} 0, \text{ если } i \neq j \\ \| Z_{\cdot j} \|^2 = n_j, \text{ если } i = j; \ n_j \text{—число элементов в } j\text{-м классе } (j\text{-й категорин фактора } z). \end{cases}$$

Коэффициент корреляционного отношения центрированного признака $x^{(i)}$ и номинальной категоризованной переменной z можно записать в виде

$$\rho^2 (x^{(i)}, z) = \frac{1}{n} \sum_{j=1}^{k} n_j \left(\overline{x}_j^{(i)} \right)^2 / s_i^2, \tag{12.8}$$

где s_i^2 — оценка дисперсии признака $x^{(i)}$; $\overline{x}_j^{(i)}$ — среднее значение признака $x^{(i)}$ для объектов, попавших в j-й класс (т. е. с j-й категорией фактора z); n_j — количество объектов в j-м классе.

Замечание. Следует помнить, что, строго говоря, в формулах (12.7) и (12.8) имеем дело с оценкой коэффициента $\rho^2 (x^{(i)}, z)$ по выборке объема n, поэтому над ними следовало бы поставить символ $\widehat{\ }$. Однако поскольку это не приводит к путанице, в данном параграфе опускаем этот символ и слово «оценка» применительно к упомянутым величинам.

Далее имеем (легко проверяется непосредственным вычислением)

$$\begin{cases} n s_i^2 = X'_{\cdot i} X_{i \cdot} = \| X_{i \cdot} \|^2; \\ n_j \overline{x}_j^{(i)} = X'_{i \cdot} Z_{\cdot j}, \end{cases}$$

где $X'_{i \cdot}$ — n-компонентный вектор, i-я строка матрицы $\mathbf{X}$.

Откуда

$$\rho^2 (x^{(i)}, z) = \sum_{j=1}^{k} \frac{1}{n_j} Z'_{\cdot j} X_{i \cdot} X'_{i \cdot} Z_{\cdot j} / \| X_{i \cdot} \|^2. \tag{12.9}$$

Используя (12.9), критерий (12.7) после некоторых преобразований можно представить в виде

$$K^2 = \sum_{j=1}^{k} \frac{1}{n_j} Z'_{\cdot j} \mathbf{U} Z_{\cdot j}, \qquad (12.10)$$

где матрица $\mathbf{U} = \sum_{i=1}^{n} \frac{v_i}{\| X_{i\cdot} \|^2} X_{i\cdot} X'_{i\cdot}$.

Учитывая известное равенство для квадратичных форм $Z'_{\cdot j} \mathbf{U} Z_{\cdot j} = \mathrm{Sp}\,[\mathbf{U}\,(Z_{\cdot j} Z'_{\cdot j})]$, критерий K^2 можно представить в более компактной форме

$$K^2 = \mathrm{Sp}\,\mathbf{U}\mathbf{R}, \qquad (12.10')$$

где $\mathbf{R} = (r_{ij},\ i = \overline{1, n},\ j = \overline{1, n}$ — матрица смежности объектов из $\mathbf{X}$, элемент $r_{lm} = 1/n$, если X_l и X_m принадлежат одному и тому же j-му классу, и 0 — в противном случае.

Элементы матрицы $n\mathbf{U}$ суть просто *взвешенные* скалярные произведения объектов (столбцов матрицы $\mathbf{X}$)

$$nu_{lm} = \sum_{i=1}^{p} v_i x_l^{(i)} x_m^{(i)} / s_i^2 = X'_l \tilde{\mathbf{V}} X_m,$$

где $\tilde{\mathbf{V}}$ — диагональная матрица, $\tilde{\mathbf{V}} = \mathrm{diag}\,(v_1/s_1^2, \ldots, v_p/s_p^2)$.

Если перейти к нормированным переменным $y = x/s_i$, то можно записать

$$nu_{lm} = \sum_{i=1}^{p} v_i y_l^{(i)} y_m^{(i)} = Y'_l \mathbf{V} Y_m.$$

Итак,

$$\mathbf{U} = \frac{1}{n} \mathbf{X} \tilde{\mathbf{V}} \mathbf{X}' = \frac{1}{n} Y \mathbf{V} Y'. \qquad (12.11)$$

С другой стороны, непосредственным вычислением легко проверяется, что

$$Z'_{\cdot j} \mathbf{U} Z_{\cdot j} = \frac{n_j}{n} \overline{X}'_j \tilde{\mathbf{V}} \overline{X_j}, \qquad (12.12)$$

где $\overline{X_j}$ — вектор средних значений для j-го класса.

Следовательно,

$$K^2 = \mathrm{Sp}\,\mathbf{V} \sum_{j=1}^{k} \frac{n_j}{n} \overline{X}_j \mathbf{D}_{\overline{X}}^{-1} \overline{X}'_j = \mathrm{Sp}\,\mathbf{V} \mathbf{D}_{\overline{X}}^{-1} \mathbf{B}_X, \qquad (12.13)$$

где $\mathbf{D}_X = \operatorname{diag}(s_1^2, \ldots, s_p^2)$, $\mathbf{B}_X$ — матрица межклассового рассеивания.

Матрицу же $\mathbf{D}_X$ можно рассматривать как полную матрицу рассеивания. Рассмотрим два случая выбора весов v_i $(i = \overline{1, p})$:

а) пусть $v_i = 1$ $(i = \overline{1, p})$ и $\tilde{v}_i = 1/s_i^2$. Тогда $\mathbf{V} = \mathbf{I}_p$ и $\tilde{V} = \mathbf{D}_X^{-1}$ и критерий примет вид

$$K^2 = \operatorname{Sp} D_X^{-1} \mathbf{B}_X = \operatorname{Sp} \mathbf{B}_Y,$$

где $\mathbf{B}_Y$ — матрица межклассового разброса для нормированных переменных. В частности, отсюда следует, что если использовать нормированные переменные, или, что то же самое, метрику вида

$$d^2(X_l, X_m) = \sum_{i=1}^{p} \frac{\left(x_l^{(i)} - x_m^{(i)}\right)^2}{s_i^2}$$

как функцию расстояния между объектами, то максимизация K^2 эквивалентна максимизации суммы корреляционных отношений между фактором z и переменными $x^{(1)}, \ldots, x^{(p)}$;

б) пусть $v_i = s_i^2$. В этом случае $\tilde{v}_i^2 = 1$ и $K^2 = \operatorname{Sp} \mathbf{B}_X$.

Рассмотрим теперь критерий T (12.5), определяемый как доля разброса, объясняемая классификацией $T = \operatorname{Sp} \mathbf{B}_X / \operatorname{Sp} \mathbf{D}_X$.

Критерий T отличается от K^2 только наличием знаменателя $\operatorname{Sp} \mathbf{D}_X$. Отсюда следует, что если в исходной метрике для получения классификации использовать критерий T, то это эквивалентно максимизации следующей *взвешенной* суммы корреляционных отношений

$$K^2 = \sum_{i=1}^{p} \frac{s_i^2}{\operatorname{Sp} \mathbf{D}_X} \rho^2(x^{(i)}, z).$$

Ясно, что если дисперсии s_i^2 сильно различаются, то получаемая классификация будет настраиваться на объяснение переменных с большими значениями s_i^2. Однозначно априорно нельзя сказать, хорошо это или плохо. Все зависит от решаемой задачи.

В табл. 12.3 суммированы результаты о соотношениях между метриками и соответствующими им критериями в терминах сумм корреляционных отношений и матриц рассеивания.

Таблица 12.3

Метрика	Формулировка критерия в терминах корреляционных отношений	Формулировка критерия в терминах матриц рассеивания
Евклидова в исходном координатном пространстве $d_{lm}^2 = \sum_{i=1}^{p} \left(x_l^{(i)} - x_m^{(i)}\right)^2$	$K^2 = \Sigma s_i^2 \rho^2 \left(x^{(i)}, z\right)$, где s_i^2 — оценка дисперсии признака $x^{(i)}$; $\rho^2 \left(x^{(i)}, z\right)$ — корреляционное отношение признака $x^{(i)}$ и z	$K^2 = \mathrm{Sp}\,\mathbf{B}_X$ или $T = \dfrac{\mathrm{Sp}\,\mathbf{B}_X}{\mathrm{Sp}\,\mathbf{D}_X}$, где $\mathbf{D}_X = \mathrm{diag}\left(s_1^2, \ldots, s_p^2\right)$; $\mathbf{B}_X$ — матрица межклассового рассеивания
Евклидова с нормированными переменными $d_{lm}^2 = \sum_{i=1}^{p} \left(x_l^{(i)} - x_m^{(i)}\right)^2 / s_i^2 = \sum_{i=1}^{p} \left(y_l^{(i)} - y_m^{(i)}\right)^2$, где $y^{(i)} = x^{(i)}/s_i$ — нормированные переменные	$K^2 = \sum_{i=1}^{p} \rho^2 \left(x^{(i)}, z\right) = \sum_{i=1}^{p} \rho^2 \left(y^{(i)}, z\right)$	$K^2 = \mathrm{Sp}\,\mathbf{D}_X^{-1}\,\mathbf{B}_X = \mathrm{Sp}\,\mathbf{B}_Y/p$, $T = K^2$
Взвешенная евклидова для исходных переменных $d_{lm}^2 = \Sigma v_i \left(x_l^{(i)} - x_m^{(i)}\right)^2$, $v_i > 0$	$K^2 = \Sigma v_i s_i^2 \rho^2 \left(x^{(i)}, z\right)$	$K^2 = \mathrm{Sp}\,\mathbf{V}\,\mathbf{B}_X$ или $T = \mathrm{Sp}\,\mathbf{V}\,\mathbf{B}_X / \mathrm{Sp}\,\mathbf{V}\mathbf{D}_X$, $\mathbf{V} = \mathrm{diag}\left(v_1, \ldots, v_p\right)$
Взвешенная евклидова с нормированными переменными $d_{lm}^2 = \sum_{i=1}^{p} v_i \left(y_l^{(i)} - y_m^{(i)}\right)^2$	$K^2 = \sum_{i=1}^{p} v_i \rho^2 \left(x^{(i)}, z\right)$	$K^2 = \mathrm{Sp}\,\mathbf{V}\mathbf{D}_X^{-1}\,\mathbf{B}_X = \mathrm{Sp}\,\mathbf{V}\mathbf{B}_Y$, $\mathbf{V} = \mathrm{diag}\left(v_1, \ldots, v_p\right)$

Классификация, объясняемая через переменные. Группировку объектов, получаемую на основе максимизации критерия (12.7), можно рассматривать как группировку, которая «объясняет» разброс переменных $x^{(1)},\ldots, x^{(p)}$ с помощью классификационного признака z. Ниже рассмотрим критерий группировки, который можно интерпретировать как критерий, «объясняющий» получаемую на основе его максимизации группировку, т. е. категории некоторой номинальной переменной z, посредством переменных $x^{(1)},\ldots, x^{(p)}$. Будет показано, что при определенном выборе метрики объясняющая группировка совпадает с объясняемой.

Введем критерий вида

$$K_1^2 = v_1 r^2(z^{(1)}, X) + \ldots + v_k r^2(z^{(k)}, X), \qquad (12.14)$$

где $r^2(z^{(i)}, X) = r_j^2$ — квадрат коэффициента множественной корреляции между фиктивной бинарной переменной $z^{(j)}$ $(j = \overline{1, k})$ и переменными $x^{(1)}, \ldots, x^{(p)}$, $v_i > 0$ — весовые коэффициенты.

Таким образом, каждая бинарная фиктивная переменная $z^{(i)}$ аппроксимируется некоторой линейной комбинацией переменных $x^{(1)}, \ldots, x^{(p)}$. Будем искать группировку (классификацию) из условия

$$Z = \arg\max K_1^2, \qquad (12.15)$$

где $Z = (z^{(1)}, \ldots, z^{(k)})'$.

Докажем следующее утверждение: если выбрать вес $v_j = 1 - n_j/n$, то критерий K_1^2 эквивалентен критерию $Q_2 = \mathrm{Sp}\mathbf{S}^{-1}\mathbf{B}$, где $\mathbf{S}$ — матрица ковариаций для X.

Для этого запишем аналитическое выражение коэффициента множественной корреляции в виде (см. § 17.2)

$$r^2(z^{(j)}, X) = \frac{1}{n} Z'_{\cdot j}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}Z_{\cdot j}/\mathbf{D}z^{(j)}$$

и

$$\mathbf{D}z^{(j)} = \frac{n_j}{n}(1 - n_j/n).$$

Матрица $\mathbf{X}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}$ является матричным представлением проекционного оператора $\mathbf{P}_X$, проектирующего n-мерные векторы на подпространство, натянутое на строки матрицы $\mathbf{X}$. С другой стороны, $\mathbf{X}\mathbf{X}' = n\mathbf{S}$, а $\mathbf{X}'Z_{\cdot j} = n_j\overline{X}_j$ — вектор средних для j-й группы. Поэтому

$$r^2(z^{(j)}, X) = \frac{n_j}{n}\overline{X}'_j\mathbf{S}^{-1}\overline{X}_j/(1 - n_j/n). \qquad (12.16)$$

Учитывая, что $\overline{X_j}\mathbf{S}^{-1}\overline{X_j} = \mathrm{Sp}\,[\mathbf{S}^{-1}(\overline{X_j}\overline{X_j})]$, получим после подстановки (12.16) в (12.14)

$$K_1^2 = Q_2 = \mathrm{Sp}\,\mathbf{S}^{-1}\mathbf{B}.$$

В отличие от критерия K^2 критерий K_1^2 афинноинвариантен. В махаланобисовой метрике $\mathbf{S} = \mathbf{I}_p$, и критерий K^2 (объясняющая группировка) и K_1^2 (объясняемая группировка) совпадают.

12.2.2. Границы значений некоторых критериев классификации. Дадим две оценки величины критерия K^2, полезные для целей интерпретации, а именно для получения представления о том, насколько удачным с формальной (критериальной) точки зрения является полученное разбиение. Эти оценки в какой-то степени заменяют статистические критерии, определяющие значимость классификации (отличие ее от случайной).

Граница снизу. Первая граница носит эвристический характер, хотя и является, по-видимому, достаточно точной и измеряет среднее значение критерия K^2 на множестве всех возможных разбиений объектов на k ($k \geqslant 2$) классов. Будем предполагать, что случайным образом многократно генерируется классификационная матрица $\mathbf{Z}$ и каждый раз вычисляется значение критерия K^2. Рассмотрим только случай нормированных переменных, полагая веса $v_i = 1$ $(i = \overline{1,p})$. Для получения оценки используем представление K^2 в виде (12.10). Значение квадратичной формы $Z'_{\cdot j}\,\mathbf{U}Z_{\cdot j}$ можно представить в виде

$$Z'_{\cdot j}\,\mathbf{U}Z_{\cdot j} = \mu \| Z_{\cdot j} \|^2 = \mu n_j,$$

где значение $\lambda_{min} < \mu < \lambda_{max}$, $\lambda_{max\,(min)}$ — соответственно максимальное (минимальное) собственное число матрицы $\mathbf{U}$.

Матрица $\mathbf{U}$ имеет не более чем p ненулевых положительных собственных чисел, совпадающих с собственными числами матрицы корреляций, и нулевое собственное число кратности не менее чем $n - p$. Средним значением собственного числа матрицы $\mathbf{U}$ будет $\lambda_{ср} = \mathrm{Sp}\,\mathbf{U}/n = p/n$. Среднее значение μ при многократном случайном выборе $Z_{\cdot j}$ будет как раз $\mu = \lambda_{ср}$.

Аналогичное равенство приближенно верно при любом j $(j = \overline{1, k})$.

Поэтому имеем приближенно

$$K^2_{ср.случ} \approx \sum_{i=1}^{k} \frac{1}{n}\,\mu_i n_i = \sum \mu_i = \frac{kp}{n},$$

где $\mu_i = \mu = \lambda_{ср}$.

Более точно

$$K^2_{\text{ср.случ}} \approx \begin{cases} \dfrac{kp}{n}, & \text{если } \dfrac{kp}{n} < 1; \\ 1, & \text{если } \dfrac{kp}{n} > 1. \end{cases} \tag{12.17}$$

Отсюда, в частности, следует, что если получена классификация Z, для которой $K^2(Z) \leqslant K^2_{\text{ср.случ}}$, то ее следует признать неудачной. Такая классификация может получиться как при неправильной настройке алгоритма кластер-анализа (например, выборе начальных центров групп), так и при отсутствии неоднородности в данных.

Граница, определяемая разбиением, предполагающим, что центры классов лежат на одной прямой. Граница $K^2_{\text{ср случ}}$ получена при усреднении значений критерия по множеству всех возможных разбиений, в том числе и очень неудачных разбиений, порожденных чисто случайным механизмом, когда точки, удаленные друг от друга, попадают в один кластер и, наоборот, очень близкие точки могут оказаться в разных кластерах. Поэтому реальное значение критерия даваемой процедурами классификации обычно существенно больше $K^2_{\text{ср случ}}$ (12.17). С другой стороны, известно, что наилучшее разбиение (в смысле любого из критериев (12.7)—(12.14) достигается в подклассе разбиений, получаемых с помощью линейных дискриминантных плоскостей [56, 60].

Естественно попробовать найти некоторое относительно просто вычисляемое линейное разбиение имеющейся матрицы данных $\mathbf{X}$. Одна из возможных кластер-процедур такого рода получается при использовании метода k-средних в предположении, что центры классов лежат на одной прямой. Получаемое таким образом значение критериальной величины обозначим через $K^2_{\min\text{ лин}}$. Очевидно, что $K^2_{\min\text{ лин}} \geqslant K^2_{\text{ср. случ}}$ и в отличие от $K^2_{\text{ср случ}}$ является величиной, зависящей от имеющейся матрицы $\mathbf{X}$, т. е. $K^2_{\min\text{ лин}} = K^2_{\min\text{ лин}}(\mathbf{X})$.

Верхняя граница для значения критерия. Для получения этой границы обратимся к представлению K^2 в виде (12.10). Векторы $Z_{\cdot 1}, \ldots, Z_{\cdot k}$ взаимно ортогональны, и норма вектора $\|Z_{\cdot j}\|^2 = n_j$. Используя экстремальные свойства собственных векторов (см. гл. 13), получаем после некоторых преобразований, что $K^2 \leqslant \sum_{i=1}^{k} \lambda_i$, где $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_k$ — собственные числа матрицы $\mathbf{U}$, упорядоченные в порядке убывания.

Так как $\sum_{i=1}^{p}\lambda_i = \sum_{i=1}^{p+l}\lambda_i = \mathrm{Sp}\mathbf{U}$ при любом $l > 0$ (поскольку имеется только p ненулевых собственных чисел), имеем

$$K^2 \leqslant \begin{cases} \sum_{i=1}^{k} \lambda_i, & \text{если } k < p; \\ \mathrm{Sp}\,\mathbf{U}, & \text{если } k > p. \end{cases} \tag{12.18}$$

Заметим далее, что ненулевые собственные числа матрицы $\mathbf{U}$ совпадают с собственными числами матрицы корреляций $\mathrm{Sp}\mathbf{U} = \mathrm{Sp}\mathbf{S}$.

12.2.3. Случай, когда центры классов лежат на одной прямой. В этом случае следует использовать метрику Махаланобиса. Расположение центров классов на одной нрямой можно рассматривать как простую модель упорядоченной классификации. Действительно, вектор средних значений i-го класса в этом случае можно представить в виде $M_i = M_0 + d_i L$ $(i = 1,k)$, где M_0 — общий центр тяжести; L — некоторый вектор, задающий направление прямой. Естественно, классы можно рассматривать упорядоченными в соответствии со значением параметра α. Если данные центрированы, то $M_0 = 0$, и, следовательно,

$$\sum_{i=1}^{k} \alpha_i d_i = 0.$$

Этот случай и будем рассматривать дальше. Без ограничения общности можно также считать, что

$$\sum_{i=1}^{k} n_i \alpha_i^2 = n.$$

Критерий для выбора вектора направления A и разбнения z запишем в виде

$$\rho^2(y,\ z) \Rightarrow \max. \tag{12.19}$$

Здесь y — новый признак, линейная комбинация исходных данных $y = a_1x_1 + \ldots + a_px_p = (A'X)$. Значение критерия (12.19) не зависит от длины вектора A. Пусть теперь классификация $\mathbf{Z}$ фиксирована. Определим вектор A, дающий критерию (12.19) максимальное значение. При этом потребуем выполнения следующего условия нормировки $A'SA = 1$ ($\mathbf{S}$ — ковариационная матрица x), т. е. будем требовать, чтобы проекция $y = A'X$ имела единичную диспер-

сию. Корреляционное отношение при выполнении условия нормировки $A'SA = 1$ можно представить в виде

$$\rho^2(y,\ z) = \sum_{j=1}^{n} \frac{1}{n_j} (A,\ \mathbf{X}'\ Z_{.j})^2. \tag{12.20}$$

После дифференцирования по A с учетом условия нормировки с помощью множителя Лагранжа получаем уравнение, которому должен удовлетворять искомый вектор A

$$(\mathbf{B} - \lambda \mathbf{S})\, A = 0, \tag{12.21}$$

где $\mathbf{B} = \sum_{j=1}^{n} \frac{n}{n_j} \overline{X}_j \overline{X}_j'$ — матрица межклассового разброса.

Это хорошо известное в дискриминантном анализе уравнение, определяющее канонический базис дискриминантного подпространства (см. гл. 19). В махаланобисовой метрике $\mathbf{S} = \mathbf{I}_p$. Используя вышесказанное, можно сформулировать следующий алгоритм направленной кластеризации.

Схема алгоритма

1. Переходим к метрике Махаланобиса и центрируем данные.

2. Задаем некоторое начальное направление $A = A^{(0)}$.

3. Производим группировку проекций объектов на A

$$z_i = (A'X_i) \ldots (i, = \overline{1,\ n}).$$

Подсчитываем центры $\overline{X_1}, \ldots, \overline{X_k}$ и матрицу $\mathbf{B}$.

Проверяем условие остановки (стабилизацию центров).

4. Пересчитываем A

$$\mathbf{B}A - \lambda \mathbf{S}A = 0,$$

здесь $\mathbf{B}$ — матрица межгруппового рассеивания по центрам.

Переходим на шаг 3.

На каждом шаге значение функционала качества не убывает, а так как он ограничен, то отсюда следует сходимость за конечное число шагов (если следить за критерием оптимизации как условием остановки).

12.3. Некоторые методические рекомендации

Использование априорной информации. Успех применения процедур классификации во многом зависит от информации, которой обладает исследователь относительно ожидаемого разделения объектов на классы. Возможно использование априорной информации в одной из следующих форм:

задание метрики в пространстве, т. е. функции расстояния между объектами (подробнее см. гл. 5, 11);

частично обучающие выборки (ЧОВ) (см. гл. 11);

неполные обучающие выборки (см. гл. 9).

Эффективность применения ОВ весьма высока. Часто ОВ суммарного объема, составляющего 5—10 % общего числа объектов, позволяют получить содержательно осмысленную классификацию, трудно реализуемую при их отсутствии.

Итеративное использование процедур классификации. Как правило, использование процедур классификации носит итеративный характер, в особенности если априорная информация отсутствует. Для получения содержательно осмысленной классификации (если она вообще потенциально возможна) полезны следующие методические приемы:

применение к данным нескольких алгоритмов классификации с последующим сравнением результатов;

применение для анализа данных нескольких метрик и нескольких вариантов параметров, управляющих работой алгоритма, с последующим сравнением результатов; при этом выбирается вариант классификации, наиболее устойчивый к вариации параметров.

Визуализация данных. Подчеркнем еще раз пользу применения средств визуализации, т. е. отображения на плоскость главных компонент и нелинейных проекций, построения гистограмм на направлениях проектирования и т. д.

Визуализация может быть использована как для выделения сгущений объектов до применения процедур классификации (тогда некоторые точки из сгущений можно попытаться использовать как ЧОВ), так и для отображения результатов работы процедуры классификации.

Результаты классификации тем устойчивее, чем больше объем выборки n и меньше соотношение p/n. В частности, поэтому полезно провести классификацию объектов, спроектированных в пространство небольшой размерности, например использовать несколько линейных или нелинейных главных компонент (см. гл. 13) и целенаправленное проецирование (см. гл. 19).

Использование дополнительных (иллюстративных) переменных. Применение иллюстративных переменных в интерпретации и оценке устойчивости разбиения описано в § 12.4.

Удаление аномальных наблюдений. Наличие аномальных наблюдений, как правило, ухудшает результаты классификации, «сжимая» имеющиеся классы. Поэтому проверка наличия таких наблюдений (см., например, § 19.5) и их удаление являются необходимым этапом перед проведением автоматической классификации.

12.4. Средства, помогающие интерпретации результатов

Предположим теперь, что в результате применения той или иной процедуры кластер-анализа или разделения смесей получена группировка исходных объектов на k групп. На дальнейшем этапе задачей исследователя является интерпретация (объяснение) полученного разделения на группы в терминах некоторого причинно-следственного механизма.

При интерпретации применяются обычно следующие средства.

Анализ состава объектов, попавших в одну группу.

Изучение статистических характеристик распределений переменных для объектов внутри каждой из групп. Для количественных переменных такими характеристиками для каждой переменной являются характеристики положения (медиана, мода, средняя величина) и рассеивания вокруг выбранной характеристики положения (обычно внутригрупповое стандартное отклонение, но может использоваться, например, и абсолютное отклонение). В качестве характеристики совместного распределения переменных внутри группы используется корреляционная матрица.

В качестве переменных-индикаторов, полезных для интерпретации группы, в первую очередь ищут такие, для которых их внутригрупповое стандартное отклонение или дисперсия много меньше стандартного отклонения (дисперсии) по всей совокупности объектов. Некоторую интерпретирующую информацию можно получить из сравнения коэффициентов корреляции между переменными для разных групп. Вспомогательным, но полезным простым средством для одновременного анализа разброса значений какой-либо переменной вокруг средних значений в каждой группе и их взаимного расположения служит линейная диаграмма. Это прямая линия, на которой расположены координаты центров групп по данной переменной с указанием интервала раз-

броса этой переменной вокруг каждого из центров (обычно $\pm \sigma$ — одно внутригрупповое стандартное отклонение).

Если среди переменных имеются неколичественные, то как индикаторы используются частоты градаций этих переменных. Если для некоторой переменной x частота ее j-й градации в i-й группе существенно выше, чем по всей выборке в среднем, то она может использоваться для интерпретации.

Использование дополнительных (иллюстративных) переменных. Кроме переменных, которые непосредственно использовались при получении классификации (активных переменных), полезно включать в рассмотрение и переменные, которые будут использованы только на этапе интерпретации. Для этих переменных в целях интерпретации оцениваются внутригрупповые статистические характеристики аналогично тому, как это делается для активных. Другое возможное их применение состоит в проведении дискриминантного анализа.

Использование дискриминантного анализа. Полученные группы объектов можно использовать как обучающие выборки для дискриминантного анализа в пространстве активных или иллюстративных переменных.

Проведение ДА в пространстве активных переменных можно использовать, с одной стороны, для целей оценки устойчивости классификации, для чего, например, подсчитывается такая характеристика, как частота ошибочной классификации (полная и попарные частоты) при применении метода скользящего экзамена. С другой стороны, для целей интерпретации можно выделить информативные переменные (пошаговый дискриминантный анализ) и использовать в интерпретации коэффициенты линейных дискриминантных функций.

Проведение ДА в пространстве иллюстративных переменных добавляет еще один аспект. Если в этом случае результаты ДА будут хорошими (низкая частота ошибок), то это будет служить дополнительным доводом в пользу предположения, что полученная группировка не случайна, а отражает некоторые существенные свойства структуры данных.

ВЫВОДЫ

1. Задача оценки качества группировки и ее интерпретации носит комплексный характер и основывается на использовании совокупности большого числа характеристик, отра-

жающих компактность групп, их взаимное расположение и распределение объектов в группах. Весьма важным, если не основным, является использование содержательных соображений.

2. В качестве средств, позволяющих оценить качество полученной группировки, полезными являются критериальные величины, характеристики компактности классов, визуальный анализ отображений на плоскости, образованные главными компонентами и факторными осями, осями, получаемыми в анализе соответствий, и особенно каноническими дискриминантными направлениями.

3. Процедуры классификации целесообразно проводить несколько раз, меняя метрики, число классов и другие параметры настройки.

4. Основной подход к интерпретации полученных групп основан на использовании статистических характеристик внутригрупповых распределений. Полезным приемом является использование дискриминантного анализа и иллюстративных переменных.

Раздел III. СНИЖЕНИЕ РАЗМЕРНОСТИ АНАЛИЗИРУЕМОГО ПРИЗНАКОВОГО ПРОСТРАНСТВА И ОТБОР НАИБОЛЕЕ ИНФОРМАТИВНЫХ ПОКАЗАТЕЛЕЙ

Глава 13. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

13.1. Сущность проблемы снижения размерности и различные методы ее решения

В исследовательской и практической статистической работе приходится сталкиваться с ситуациями, когда общее число p признаков $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, регистрируемых на каждом из множества обследуемых объектов (стран, городов, предприятий, семей, пациентов, технических или экологических систем), очень велико — порядка ста и более. Тем не менее имеющиеся многомерные наблюдения

$$X_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(p)} \end{pmatrix}, \quad i = 1, 2, \ldots, n, \tag{13.1}$$

следует подвергнуть статистической обработке, осмыслить либо ввести в базу данных для того, чтобы иметь возможность их использовать в нужный момент.

Желание статистика представить каждое из наблюдений (13.1) в виде вектора Z некоторых *вспомогательных* показателей $z^{(1)}$, $z^{(2)}$, ..., $z^{(p')}$ с существенно меньшим (чем p) числом компонент p' бывает обусловлено в первую очередь следующими причинами:

необходимостью *наглядного представления* (визуализации) исходных данных (13.1), что достигается их проецированием на специально подобранное трехмерное пространство ($p' = 3$), плоскость ($p' = 2$) или числовую прямую (задачам такого типа посвящен раздел IV);

стремлением к *лаконизму исследуемых моделей*, обусловленному необходимостью упрощения счета и интерпретации полученных статистических выводов;

необходимостью существенного *сжатия объемов хранимой статистической информации* (без видимых потерь в ее информативности), если речь идет о записи и хранении массивов типа (13.1) в специальной базе данных.

При этом новые (вспомогательные) признаки $z^{(1)}$, $z^{(2)}$, ..., $z^{(p')}$ могут выбираться из числа исходных или определяться по какому-либо правилу по совокупности исходных признаков, например как их линейные комбинации. При формировании новой системы признаков к последним предъявляются разного рода требования, такие, как наибольшая информативность (в определенном смысле), взаимная некоррелированность, наименьшее искажение геометрической структуры множества исходных данных и т. п. В зависимости от варианта формальной конкретизации этих требований (см. ниже, а также раздел IV) приходим к тому или иному алгоритму снижения размерности. Имеется, по крайней мере, три основных типа принципиальных предпосылок, обусловливающих возможность перехода от большого числа p исходных показателей состояния (поведения, эффективности функционирования) анализируемой системы к существенно меньшему числу p' наиболее информативных переменных. Это, во-первых, *дублирование информации, доставляемой сильно взаимосвязанными признаками*; во-вторых, *неинформативность признаков, мало меняющихся при переходе от одного объекта к другому* (малая «вариабельность» признаков); в-третьих, *возможность агрегирования*, т. е. простого или «взвешенного» суммирования, по некоторым признакам.

Формально задача перехода (с наименьшими потерями в информативности) к новому набору признаков $\tilde{z}^{(1)}$, $\tilde{z}^{(2)}$, ..., $\tilde{z}^{(p')}$ может быть описана следующим образом. Пусть $Z = Z(X)$ — некоторая p-мерная вектор-функция исходных переменных $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ ($p' \ll p$) и пусть $I_{p'}(Z(X))$ — определенным образом заданная мера информативности p'-мерной системы признаков $Z(X) = (z^{(1)}(X), ..., z^{(p')}(X))$. Конкретный выбор функционала $I_{p'}(Z)$ зависит от специфики решаемой реальной задачи и опирается на один из возможных критериев: критерий *автоинформативности*, нацеленный на максимальное сохранение информации, содержащейся в исходном массиве $\{X_i\}_{i=\overline{1,n}}$ относительно самих исходных признаков; и критерий *внешней информативности*, нацеленный на максимальное «выжимание» из $\{X_i\}_{i=\overline{1,n}}$ информации, содержащейся в этом массиве относительно некоторых других (внешних) показателей.

Задача заключается в определении такого набора признаков $\tilde{Z}$, найденного в классе $\mathbf{F}$ допустимых преобразований исходных показателей $x^{(1)}$, ..., $x^{(p)}$, что

$$I_{p'}\left(\tilde{Z}(X)\right) = \max_{Z \in \mathbf{F}} \{I_{p'}(Z(X))\}. \qquad (13.2)$$

Тот или иной вариант конкретизации этой постановки (определяющий конкретный выбор меры информативности $I_{p'}(Z)$ и класса допустимых преобразований) приводит к конкретному методу снижения размерности: к методу главных компонент, факторному анализу, экстремальной группировке параметров и т. д.

Поясним это на примерах.

13.1.1. Метод главных компонент (см. § 13.2—§ 13.6). Именно к p' первым главным компонентам придет исследователь, если в качестве класса допустимых преобразований F определит всевозможные линейные ортогональные нормированные комбинации исходных показателей, т. е.

$$z^{(j)}(X) = c_{j1}(x^{(1)} - \mu^{(1)}) + \ldots + c_{jp}(x^{(p)} - \mu^{(p)});$$

$$\sum_{\nu=1}^{p} c_{j\nu}^2 = 1, \quad j = 1, 2, \ldots, p; \tag{13.3}$$

$$\sum_{\nu=1}^{p} c_{j\nu} c_{k\nu} = 0, \quad j, k = 1, 2, \ldots, p; \quad j \neq k$$

(здесь $\mu^{(\nu)} = \mathbf{E}x^{(\nu)}$ — математическое ожидание $x^{(\nu)}$), а в качестве меры информативности p'-мерной системы показателей $(z^{(1)}(X), \ldots, z^{(p')}(X))$ выражение

$$I_{p'}(Z(X)) = \frac{\mathbf{D}z^{(1)} + \ldots + \mathbf{D}z^{(p')}}{\mathbf{D}x^{(1)} + \ldots + \mathbf{D}x^{(p)}} \tag{13.4}$$

(здесь D, как и ранее, знак операции вычисления дисперсии соответствующей случайной величины).

13.1.2. Факторный анализ (см. гл. 14). Как известно (см. § 14.1), модель факторного анализа объясняет структуру связей между исходными показателями $x^{(1)}, \ldots, x^{(p)}$ тем, что поведение каждого из них статистически зависит от одного и того же набора так называемых *общих факторов* $y^{(1)}, \ldots, y^{(p')}$, т. е.

$$x^{(j)} - \mu^{(j)} = \sum_{\nu=1}^{p'} q_{j\nu} y^{(\nu)} + u^{(j)} \quad (j = 1, 2, \ldots, p),$$

где $q_{j\nu}$ — «нагрузка» общего фактора $y^{(\nu)}$ на исходный показатель $x^{(j)}$, а $u^{(j)}$ — остаточная «специфическая» случайная компонента, причем $\mathbf{E}y^{(\nu)} = 0$, $\mathbf{E}u^{(j)} = 0$, $\mathbf{D}y^{(\nu)} = 1$ и $y^{(1)}, \ldots, y^{(p)'}, u^{(1)}, \ldots, u^{(p)}$ — попарно некоррелированы.

Оказывается, если F определить как класс всевозможных линейных комбинаций $x^{(1)}, \ldots, x^{(p)}$ с учетом упомянутых ог-

ранжений на $y^{(v)}$, а в качестве меры информативности p-мерной системы показателей выбрать величину $I_{p'}(Z(X)) = 1 - \|\mathbf{R}_X - \mathbf{R}_{\widehat{X}}\|^2$, то решение оптимизационной задачи (13.2) совпадает с вектором общих факторов $(y^{(1)}, \ldots, y^{(p')})$ в модели факторного анализа. Здесь $\mathbf{R}_X$ — корреляционная матрица исходных показателей $x^{(1)}, \ldots, x^{(p)}$, $\mathbf{R}_{\widehat{X}}$ — корреляционная матрица показателей $\widehat{x^{(j)}} = \sum_{v=1}^{p'} q_{jv}\, y^{(v)}$, а $\|A\|$ — евклидова норма матрицы $\mathbf{A}$.

13.1.3. Метод экстремальной группировки признаков (см. п. 14.2.1). В данном методе речь идет о таком разбиении совокупности исходных показателей $x^{(1)}, \ldots, x^{(p)}$ на заданное число p' групп $S_1, \ldots, S_{p'}$, что признаки, принадлежащие одной группе, были бы взаимокоррелированы сравнительно сильно, в то время как признаки, принадлежащие к разным группам, были бы коррелированы слабо. Одновременно решается задача замены каждой (i-й) группы сильно взаимокоррелированных исходных показателей одним вспомогательным «равнодействующим» показателем $z^{(i)}$, который, естественно, должен быть в тесной корреляционной связи с признаками своей группы. Определив в качестве класса допустимых преобразований $\mathbf{F}$ исходных показателей все нормированные ($\mathbf{D}z^{(i)} = 1$) линейные комбинации $x^{(1)}, \ldots, x^{(p)}$, ищем решение $(S_1^*, \ldots, S_{p'}^*; \widetilde{z}^{(1)}, \ldots, \widetilde{z}^{(p')})$, максимизируя (по S и $Z(X)$) функционал

$$I_{p'}(Z(X);\ S) = \sum_{x^{(k)} \in S_1} r^2(x^{(k)},\ z^{(1)}) + \ldots +$$

$$+ \sum_{x^{(k)} \in S_{p'}} r^2(x^{(k)},\ z^{(p')}),$$

где $r(x, z)$ — коэффициент корреляции между переменными x и z.

13.1.4. Многомерное шкалирование (см. гл. 16). В ряде ситуаций и в первую очередь в ситуациях, когда исходные статистические данные получают с помощью специальных опросов, анкет, экспертных оценок, возможны случаи, когда элементом первичного наблюдения является не состояние i-го объекта, описываемого вектором X_i, а характеристика ρ_{ij} попарной близости (отдаленности) двух объектов (или признаков) соответственно с номерами i и j.

В этом случае исследователь располагает в качестве массива исходных статистических данных матрицей размера $n \times n$ (если рассматриваются характеристики попарной бли-

зости *объектов*) или $p \times p$ (если рассматриваются характеристики попарной близости *признаков*) вида

$$\rho = (\rho_{ij}),\ i,\ j = 1,\ 2,\ \ldots,\ m,\ m = n \text{ или } m = p, \qquad (13.5)$$

где величины ρ_{ij} интерпретируются либо как расстояния между объектами (признаками) i и j, либо как ранги, задающие упорядочение этих расстояний. Задача многомерного шкалирования состоит в том, чтобы «погрузить» наши объекты (признаки) в такое p'-мерное пространство ($p' \ll \min(p, n)$), т. е. так выбрать координатные оси $0z^{(1)}, \ldots, 0z^{(p')}$, чтобы исходная геометрическая конфигурация совокупности анализируемых точек-объектов (или точек-признаков), заданных с помощью (13.1) или (13.5), оказалась бы наименее искаженной в смысле некоторого критерия средней «степени искажения» $\Delta(Z)$ взаимных попарных расстояний.

Одна из достаточно общих схем многомерного шкалирования определяется критерием

$$\Delta(Z) = \sum_{i,j=1}^{n} d_{ij}^{\alpha} \left| \widehat{d}_{ij}(Z) - d_{ij} \right|^{\beta},$$

где d_{ij} — расстояние между объектами O_i и O_j в *исходном* пространстве, $\widehat{d}_{ij}(Z)$ — расстояние между теми же объектами в *искомом* пространстве меньшей размерности p', а α и β — свободные параметры, выбор конкретных значений которых производится по усмотрению исследователя.

Определив меру информативности искомого набора признаков Z, например, как величину, обратную упомянутой выше величине степени искажения геометрической структуры исходной совокупности точек, сведем эту задачу к общей постановке (13.2), полагая

$$I_{p'}(Z) = \left[1 + \sum_{i,j=1}^{n} d_{ij}^{\alpha} \left| \widehat{d}_{ij}(Z) - d_{ij} \right|^{\beta} \right]^{-1}.$$

13.1.5. Отбор наиболее информативных показателей в моделях дискриминантного анализа (см. § 1.4; 2.5). Приведенные выше функционалы являются измерителями *автоинформативности* соответствующей системы признаков. Приведем теперь примеры критериев *внешней информативности*. В частности, нас будет интересовать информативность системы показателей $z^{(1)}(X), \ldots, z^{(p')}(X)$ с точки зрения правильности классификации объектов по этим показателям в схеме дискриминантного анализа. При этом класс допустимых преобразований $\mathbf{F}$ определим исходя из требований, что в качестве $z^{(k)}(X)$ могут рассматриваться лишь представи-

тели набора исходных показателей, т. е. $Z(X) = (x^{(i_1)}, x^{(i_2)}, \ldots, x^{(i_{p'})})$. Распространенным исходным тезисом при решении задачи выявления наиболее информативных p' показателей из исходного набора $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ является утверждение, что вектор показателей $(x^{(i_1)}, x^{(i_2)}, \ldots, x^{(i_{p'})})$ заданной размерности p' тем более информативен, чем больше различие в законах его вероятностного распределения, определенных в разных классах в рассматриваемой задаче классификации. Если ввести меру попариого различия $\delta\{P_i(Z), P_j(Z)\}$ законов $P_1(Z), \ldots, P_k(Z)$, описывающих распределение вероятностей вектора признаков $Z = (x^{(i_1)}, \ldots, x^{(i_{p'})})$ в классах с номерами 1, 2, ..., k, то можно формализовать вышеприведенный принцип отбора наиболее информативных показателей $x^{(i_1)}, \ldots, x^{(i_{p'})}$, определяя их из условия максимизации (по $i_1, i_2, \ldots, i_{p'}$) величины

$$I_{p'}(Z(X)) = \sum_{i,j=1}^{k} \delta\{P_i(Z), P_j(Z)\}.$$

Наиболее употребительные меры различия между законами распределения вероятностей $\delta\{P_i, P_j\}$ — это расстояние информационного типа (расстояние Кульбака, расстояние Махаланобиса), а также «расстояние по вариации» (подробнее об этом см. в [154, с. 76—84].

13.1.6. Отбор наиболее информативных переменных в моделях регрессии (см. [12, гл. 8]). При построении зависимостей регрессионного типа одним из центральных оказывается вопрос выявления сравнительно небольшого числа p' переменных (из априорного набора $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$), наиболее существенно влияющих на поведение исследуемого результирующего признака y.

Таким образом, как и в предыдущем пункте, класс **F** состоит из всевозможных наборов переменных $Z = (x^{(i_1)}, \ldots, x^{(i_{p'})})$, отобранных из исходного множества факторов-аргументов $x^{(1)}, \ldots, x^{(p)}$, и имеем дело с критерием *внешней информативности* таких наборов. Его вид обычно задается с помощью множественного коэффициента детерминации $R^2_y(x^{(i_1)}, \ldots, x^{(i_{p'})})$ — характеристики степени тесноты связи показателя y с набором переменных $x^{(i_1)}, \ldots, x^{(i_{p'})}$. При этом для фиксированной размерности $p' < p$ набор переменных $x^{(i_1^*)}, \ldots, x^{(i_{p'}^*)}$ будет, очевидно, считаться наиболее информативным (с точки зрения точности описания поведения показателя y), если значение меры информативности на этом наборе достигает максимума.

13.1.7. Сведéние нескольких частных критериальных показателей к единому интегральному (см. гл. 15). Речь идет о ситуациях, в которых «качество функционирования» исследуемой системы или объекта (предприятия, сложного изделия, отдельного специалиста и т.д.) характеризуется *набором* поддающихся измерению частных критериальных показателей $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$. Однако требуется перейти к некоторой не поддающейся непосредственному измерению *скалярной интегральной* оценке y. При этом постулируется, что латентный показатель y является функцией *известного* общего вида от $x^{(1)}, \ldots, x^{(p)}$, т. е. $y = f(x^{(1)}, \ldots, x^{(p)}; \Theta)$, и требуется подобрать лишь неизвестное значение параметра (вообще говоря, векторного) Θ.

Для решения этой задачи к зарегистрированной в результате контрольного обследования исходной статистической информации вида (13.1) приходится добавлять один из следующих вариантов *экспертной* информации о показателе y.

Вариант 1: *балльная оценка* «выходного качества» y, т. е. значения $y_{1э}, y_{2э}, \ldots, y_{nэ}$, экспертно оценивающие в определенной балльной шкале «выходное качество» 1-го, 2-го, ..., n-го объектов.

Вариант 2: *ранжирование* анализируемых объектов, т. е. их упорядочение по степени убывания «выходного качества» y; таким образом будем иметь ранги $R_э = \{R_{iэ}\}_{i=\overline{1,n}}$, т. е. порядковые номера объектов в этом упорядоченном ряду.

Вариант 3: результаты *попарных сравнений* анализируемых объектов по интересующему нас «выходному качеству» или результат разбиения контрольной совокупности объектов на группы, однородные с точки зрения «выходного качества»; и в том и в другом случае экспертные данные могут быть представлены с помощью булевой матрицы $\Gamma = (\gamma_{ij_э})_{i,j=\overline{1,n}}$, где $\gamma_{ij_э} = 1$, если O_i не хуже O_j, $\gamma_{ij} = 0$ в противном случае.

Алгоритмы определения неизвестного параметра Θ используют в качестве исходной статистическую информацию (13.1), дополненную одним из вариантов экспертной информации (поэтому метод называется экспертно-статистическим), и построены на следующей идее. Если было бы известно значение параметра $\widehat{\Theta}$, можно было бы вычислить значение целевой функции $f(x^{(1)}, \ldots, x^{(p)}; \widehat{\Theta})$ для каждого из контрольных объектов и определить с помощью этой целевой функции и балльные оценки $f(x^{(1)}, \ldots, x^{(p)}; \widehat{\Theta})$, и ран-

ги $R(\widehat{\Theta}) = \{R_i(\widehat{\Theta})\}_{i=\overline{1,n}}$, и матрицу парных сравнений $\Gamma(\widehat{\Theta}) = (\gamma_{ij}(\widehat{\Theta}))$, $i, j = \overline{1, n}$.

Поэтому если хотим формализовать с помощью целевой функции $f(X; \Theta)$ экспертные критерийные установки, в соответствии с которыми формируется единый интегральный показатель «выходного качества» y, естественно подчинить алгоритм поиска параметра Θ оптимизационному критерию вида

$$I_1(Z(X, \Theta)) =$$

$$= \begin{cases} \left[1 + \sum_{i=1}^{n} \left(y_{i_3} - f\left(x_i^{(1)}, \ldots, x_i^{(p)}; \Theta\right)\right)^2\right]^{-1} & \text{в варианте 1;} \\ r(R_3, R(\Theta)) & \text{в варианте 2;} \\ \left[1 + \sum_{i=1}^{n} |\gamma_{ij_3} - \widehat{\gamma}_{ij}(\Theta)|\right]^{-1} & \text{в варианте 3} \end{cases}$$

(здесь под $r(S, Q)$ подразумевается коэффициент ранговой корреляции Спирмэна между ранжировками S и Q). Разработаны алгоритмы и программы, позволяющие вычислять Θ в задаче максимизации критерия $I_1(Z(X; \Theta))$ для всех трех вариантов (см. гл. 15).

13.2. Определение, вычисление и основные числовые характеристики главных компонент

Во многих задачах обработки многомерных наблюдений и, в частности, в задачах классификации исследователя интересуют в первую очередь лишь те признаки, которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного объекта к другому.

С другой стороны, не обязательно для описания состояния объекта использовать какие-то из исходных, непосредственно замеренных на нем признаков. Так, например, для определения специфики фигуры человека при покупке одежды достаточно назвать значения двух признаков (размер — рост), являющихся производными от измерений ряда параметров фигуры. При этом, конечно, теряется какая-то доля информации (портной измеряет до одиннадцати параметров на клиенте), как бы огрубляются (при агрегировании) получающиеся при этом классы. Однако, как показали исследо-

вания, к вполне удовлетворительной классификации людей с точки зрения специфики их фигуры приводит система, использующая три признака, каждый из которых является некоторой комбинацией от большого числа непосредственно замеряемых на объекте параметров.

Именно эти принципиальные установки заложены в сущность того линейного преобразования исходной системы признаков, которое приводит к главным компонентам. Формализуются же эти установки следующим образом.

Следуя общей оптимизационной постановке задачи снижения размерности (13.2) и полагая анализируемый признак X p-мерной случайной величиной с вектором средних значений $\boldsymbol{\mu} = (\mu^{(1)}, \ldots, \mu^{(p)})$ и ковариационной матрицей $\boldsymbol{\Sigma} = (\sigma_{ij})$ $(i, j = 1, 2, \ldots, p)$, вообще говоря, неизвестными, определим меру (критерий) информативности $Ip'(Z)$ вспомогательной p'-мерной системы показателей $Z = (z^{(1)}, \ldots, z^{(p')})$ с помощью (13.4), а класс допустимых преобразований — в виде (13.3). Тогда при любом фиксированном $p' = 1, 2, \ldots, p$ вектор искомых вспомогательных переменных $\widetilde{Z}(X) = (\widetilde{z}^{(1)}(X), \ldots, \widetilde{z}^{(p')}(X))'$ определяется как такая линейная комбинация

$$\widetilde{Z} = \mathbf{L}X$$

(где матрица

$$\mathbf{L} = \begin{pmatrix} l_{11} & \ldots & l_{1p} \\ \ldots & \ldots & \ldots \\ l_{p'1} & \ldots & l_{p'p} \end{pmatrix}, \tag{13.6}$$

а ее строки удовлетворяют условию ортогональности), что

$$I_{p'}\left(\widetilde{z}^{(1)}(X), \ldots, \widetilde{z}^{(p')}(X)\right) = \max_{Z(X) \in \mathbf{F}} I_{p'}(Z(X)).$$

Полученные таким образом переменные $\widetilde{z}^{(1)}(X), \ldots, \widetilde{z}^{(p)}(X)$ и называют главными компонентами вектора X. Поэтому можно дать следующее определение главных компонент.

Первой главной компонентой $\widetilde{z}^{(1)}(X)$ исследуемой системы показателей $X = (x^{(1)}, \ldots, x^{(p)})'$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая среди всех прочих нормированно-центрированных линейных комбинаций переменных $x^{(1)}, \ldots, x^{(p)}$ обладает наибольшей дисперсией.

k-й главной компонентой $(k = 2, 3, \ldots, p)$ исследуемой системы показателей $X = (x^{(1)}, \ldots x^{(p)})'$ называется такая

нормированно-центрированная линейная комбинация этих показателей, которая не коррелирована с $k-1$ предыдущими главными компонентами и среди всех прочих нормированно-центрированных и не коррелированных с предыдущими $k-1$ главными компонентами линейных комбинаций переменных $x^{(1)}, \ldots, x^{(p)}$ обладает наибольшей дисперсией.

Замечание 1 (переход к центрированным переменным). Поскольку, как увидим ниже, решение задачи (а именно вид матрицы линейного преобразования $\mathbf{L}$) зависит только от элементов ковариационной матрицы $\boldsymbol{\Sigma}$, которые в свою очередь не изменяются при замене исходных переменных $x^{(j)}$ переменными $x^{(j)} - c^{(j)}$ ($c^{(j)}$ — произвольные постоянные числа), то в дальнейшем будем считать, что исходная система показателей уже *центрирована*, т. е. что $\mathrm{E}x^{(j)} = 0$, $j = 1,2, \ldots, p$. В статистической практике этого добиваются, переходя к наблюдениям $\overset{\circ}{x}{}_i^{(j)} = x_i^{(j)} - \overline{x}^{(j)}$, где $\overline{x}^{(j)} = \sum_{i=1}^{n} x_i^{(j)}/n$ (для упрощения обозначений волнистую черту над центрированной переменной и над главной компонентой в дальнейшем ставить не будем).

Замечание 2 (переход к выборочному варианту). Поскольку в реальных статистических задачах располагаем *лишь оценками* $\widehat{\boldsymbol{\mu}}$ и $\widehat{\boldsymbol{\Sigma}}$ соответственно вектора средних $\boldsymbol{\mu}$ и ковариационной матрицы $\boldsymbol{\Sigma}$, то во всех дальнейших рассуждениях под $\boldsymbol{\mu}^{(j)}$ понимается $\overline{x}^{(j)}$, а под σ_{kj} — выборочная ковариация $\widehat{\sigma}_{kj} = \sum_{i=1}^{n} (x_i^{(k)} - \overline{x}^{(k)})(x_i^{(j)} - \overline{x}^{(j)})/n$ ($j, k = 1, 2, \ldots, p$).

Вычисление главных компонент. Из определения главных компонент следует, что для вычисления первой главной компоненты необходимо решить оптимизационную задачу вида

$$\begin{cases} \mathbf{D}(l_1 X) \to \max\limits_{l_1}; \\ l_1 l_1' = 1, \end{cases} \tag{13.7}$$

где l_1 — первая строка матрицы $\mathbf{L}$ (см. (13.6)). Учитывая центрированность переменной X (т. е. $\mathbf{E}X = \mathbf{0}$) и то, что $\mathbf{E}(XX') = \boldsymbol{\Sigma}$, имеем

$$\mathbf{D}(l_1 X) = \mathbf{E}(l_1 X)^2 = \mathbf{E}(l_1 XX' l_1') = l_1 \boldsymbol{\Sigma} l_1'.$$

Следовательно, задача (13.7) может быть записана

$$\begin{cases} l_1 \Sigma l_1' \to \max\limits_{l_1}; \\ l_1 l_1' = 1. \end{cases} \tag{13.7'}$$

Вводя функцию Лагранжа $\varphi(l_1, \lambda) = l_1 \Sigma l_1' - \lambda (l_1 l_1' - 1)$ и дифференцируя ее по компонентам вектор-столбца l_1', имеем

$$\frac{\partial \varphi}{\partial l_1'} = 2\Sigma l_1' - 2\lambda l_1',$$

что дает систему уравнений для определения l_1:

$$(\Sigma - \lambda \mathbf{I}) l_1' = \mathbf{0} \tag{13.8}$$

(здесь $\mathbf{0} = (0, 0, \ldots 0)'$ — p-мерный вектор-столбец из нулей).

Для того чтобы существовало ненулевое решение системы (13.8) (а оно должно быть ненулевым, так как $l_1 l_1' = 1$), матрица $\Sigma - \lambda \mathbf{I}$ должна быть *вырожденной*, т. е.

$$|\Sigma - \lambda \mathbf{I}| = 0. \tag{13.9}$$

Этого добиваются за счет подбора соответствующего значения λ. Уравнение (13.9) (относительно λ) называется *характеристическим* для матрицы Σ. Известно, что при симметричности и неотрицательной определенности матрицы Σ (каковой она и является как всякая ковариационная матрица) это уравнение имеет p вещественных неотрицательных корней $\lambda_1 \geqslant \lambda_2 \ldots \geqslant \lambda_p \geqslant 0$, называемых *характеристическими* (или *собственными*) *значениями* матрицы Σ.

Учитывая, что $\mathrm{D}z^{(1)} = \mathbf{D}(l_1 X) = l_1 \Sigma l_1'$ (см. выше) и $l_1 \Sigma l_1' = \lambda$ (последнее соотношение следует из (13.8) после его умножения слева на l_1, с учетом $l_1 l_1' = 1$), получаем

$$\mathbf{D}z^{(1)}(X) = \lambda.$$

Поэтому для обеспечения максимальной величины дисперсии переменной $z^{(1)}$ нужно выбрать из p собственных значений матрицы Σ *наибольшее*, т. е.

$$\mathbf{D}z^{(1)}(X) = \lambda_1.$$

Подставляем λ_1 в систему уравнений (13.8) и, решая ее относительно $l_{11}, \ldots, l_{1p}$, определяем компоненты вектора l_1.

Таким образом, *первая главная компонента получается как линейная комбинация* $z^{(1)}(X) = l_1 \cdot X$, где l_1 — *собственный вектор матрицы* Σ, *соответствующий наибольшему собственному числу этой матрицы.*

Далее аналогично можно показать, что $z^{(k)}(X) = l_k \cdot X$, где l_k — собственный вектор матрицы Σ, соответствующий k-му по величине собственному значению λ_k этой матрицы.

Таким образом соотношения для определения всех p главных компонент вектора X могут быть представлены в виде

$$Z = \mathbf{L}X,$$

где $Z = (z^{(1)}, \ldots, z^{(p)})'$, $X = (x^{(1)}, \ldots, x^{(p)})'$, а матрица $\mathbf{L}$ состоит из строк $l_j = (l_{j1}, \ldots, l_{jp})$, $j = \overline{1, p}$, являющихся собственными векторами матрицы Σ, соответствующими собственным числам λ_j. При этом сама матрица $\mathbf{L}$ по построению является *ортогональной*, т. е.

$$\mathbf{L}\mathbf{L}' = \mathbf{L}'\mathbf{L} = \mathbf{I}.$$

Основные числовые характеристики главных компонент. Определим основные числовые характеристики (средние значения, дисперсии, ковариации) главных компонент в терминах основных числовых характеристик исходных переменных и собственных значений матрицы Σ:

а) $\mathbf{E}Z = \mathbf{E}(\mathbf{L}X) = \mathbf{L} \cdot \mathbf{E}X = 0$;

б) ковариационная матрица вектора главных компонент:

$$\Sigma_Z = \mathbf{E}(ZZ') = \mathbf{E}((\mathbf{L}X)(\mathbf{L}X)') = \mathbf{E}(\mathbf{L}XX'\mathbf{L}') =$$
$$= \mathbf{L} \cdot \mathbf{E}(XX') \cdot \mathbf{L}' = \mathbf{L} \cdot \Sigma \cdot \mathbf{L}'.$$

Умножая слева соотношения

$$(\Sigma - \lambda_k \mathbf{I}) l_k' = 0 \quad (k = \overline{1, p})$$

на l_j $(j = \overline{1, p})$, получаем, что

$$\mathbf{L}\Sigma\mathbf{L}' = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots \\ 0 & & \lambda_p \end{pmatrix}$$

и, следовательно:

$$\Sigma_z = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots \\ 0 & & \lambda_p \end{pmatrix} \tag{13.10}$$

Из (13.10), в частности, следует подтверждение взаимной некоррелированности главных компонент, а также $\mathbf{D}z^{(k)} = \lambda_k$ $(k=\overline{1,\ p})$,

в) сумма дисперсий исходных признаков равна сумме дисперсий всех главных компонент. Действительно, $\sum_{k=1}^{p} \mathbf{D}z^{(k)} = \mathrm{Sp}\,\mathbf{\Sigma}_Z = \mathrm{Sp}\,(\mathbf{L}\mathbf{\Sigma}\mathbf{L}') = \mathrm{Sp}((\mathbf{L}\mathbf{\Sigma})\cdot\mathbf{L}') = \mathrm{Sp}\,(\mathbf{L}'\cdot(\mathbf{L}\mathbf{\Sigma})) = \mathrm{Sp}((\mathbf{L}'\mathbf{L})\mathbf{\Sigma}) = \mathrm{Sp}\mathbf{\Sigma} = \sum_{k=1}^{p} \mathbf{D}X^{(k)}$;

г) обобщенная дисперсия исходных признаков (X) равна обобщенной дисперсии главных компонент (Z). Действительно, обобщенная дисперсия вектора Z равна

$$|\mathbf{\Sigma}_Z| = |\mathbf{L}\mathbf{\Sigma}\mathbf{L}'| = |(\mathbf{L}\mathbf{\Sigma})\mathbf{L}'| = |\mathbf{L}'(\mathbf{L}\mathbf{\Sigma})| = |(\mathbf{L}\mathbf{L}')\mathbf{\Sigma}| = |\mathbf{\Sigma}|.$$

Следствие. Из б) и в), в частности, следует, что критерий информативности метода главных компонент (13.9) может быть представлен в виде

$$I_{p'}(Z(X)) = \frac{\lambda_1 + \ldots + \lambda_{p'}}{\lambda_1 + \ldots + \lambda_p}, \qquad (13.9')$$

где $\lambda_1, \lambda_2, \ldots, \lambda_p$ — собственные числа ковариационной матрицы $\mathbf{\Sigma}$ вектора X, расположенные в порядке убывания.

Кстати, представление $I_{p'}(Z(X))$ в виде (13.9′) дает исследователю некоторую основу, опорную точку зрения, при вынесении решения о том, сколько последних главных компонент можно без особого ущерба изъять из рассмотрения, сократив тем самым размерность исследуемого пространства.

Действительно, анализируя с помощью (13.9′) изменение относительной доли дисперсии, вносимой первыми p' главными компонентами, в зависимости от числа этих компонент, можно разумно определить число компонент, которое целесообразно оставить в рассмотрении. Так, при изменении $I_{p'}$, изображенном на рис. 13.1, очевидно, целесообразно было бы сократить размерность пространства с $p = 10$ до $p' = 3$, так как добавление всех остальных семи главных компонент может повысить суммарную характеристику рассеяния не более чем на 10 %.

Замечание 3. Использование главных компонент оказывается наиболее естественным и плодотворным в ситуациях, в которых все компоненты $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ исследуемого вектора X имеют *общую физическую природу* и соответственно *измерены в одних и тех же единицах*. К таким примерам можно отнести исследование структуры бюд-

жета времени индивидуумов (все $x^{(i)}$ измеряются в единицах времени), исследование структуры потребления семей (все $x^{(i)}$ измеряются в денежных единицах), исследование общего развития и умственных способностей индивидуумов с помощью специальных тестов (все $x^{(i)}$ измеряются в баллах), разного рода антропологические исследования (все $x^{(i)}$ измеряются в единицах меры длины) и т.д. Если же различные признаки $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ измеряются в различных

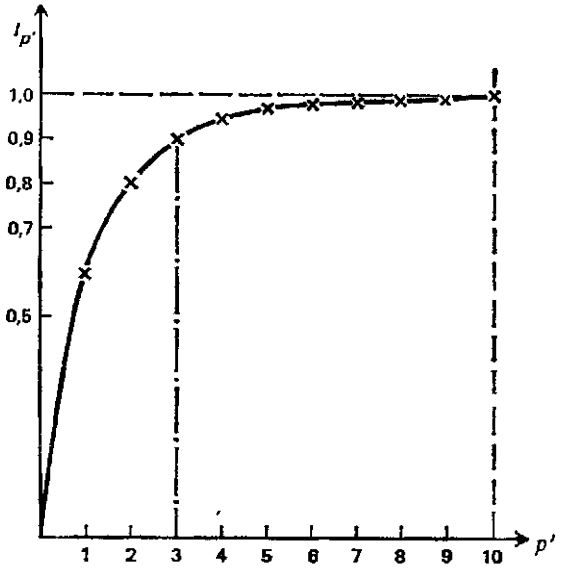


Рис. 13.1. Измененне относительной долн суммарной дисперсин нсследуемых признаков, обусловленной первымн p' главнымн компонентами, в зависнмости от p' (случай $p=10$)

единицах, то результаты исследования с помощью главных компонент будут существенно зависеть от выбора масштаба и природы единиц измерения. Поэтому в подобных ситуациях исследователь предварительно переходит к вспомогательным безразмерным признакам $x^{*(i)}$, например с помощью нормирующего преобразования

$$x_{\nu}^{*(i)} = \frac{x_{\nu}^{(i)}}{\sqrt{\hat{\sigma}_{ii}}} \quad \begin{pmatrix} i = 1, 2, \ldots, p \\ \nu = 1, 2, \ldots, n \end{pmatrix},$$

где σ_{ii} соответствует ранее введенным обозначениям, а затем строит главные компоненты относительно этих вспомо-

гательных признаков X^* и их ковариационной матрицы $\widehat{\Sigma}_{X^*}$, которая, как легко видеть, является одновременно выборочной корреляционной матрицей R исходных наблюдений X_i.

Замечание 4. В некоторых задачах оказывается полезным понятие так называемых *обобщенных* главных компонент, при определении которых оговаривают более общие (чем $\sum_{j=1}^{p} l_{ij}^2 = 1$) ограничения на коэффициенты l_{ij}, т. е. требуют, чтобы

$$\sum_{k=1}^{p}\sum_{j=1}^{p} l_{ij}\omega_{kj}l_{ik} = 1,$$

где ω_{kj} — некоторые дополнительно введенные веса. Очевидно, если $\omega_{kj} = 1$ при $k = j$ и $\omega_{kj} = 0$ при $k \neq j$, то имеем обычное условие нормировки коэффициентов l_{ij} и обычные главные компоненты. Можно показать, что при такой модификации условий нормировки коэффициенты $l_i = (l_{i1}, l_{i2}, \ldots, l_{ip})$, с помощью которых обобщенные главные компоненты $z^{(i)}$ выражаются через исходные признаки $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, определяются как решения уравнений

$$(\Sigma - \tilde{\lambda}_i \Omega) l_i' = 0,$$

где $\tilde{\lambda}_i$ — i-й по величине корень уравнения $|\Sigma - \tilde{\lambda}\Omega| = 0$, а матрица $\Omega = (\omega_{ij})$, $i, j = 1, 2, \ldots, p$ — некоторая положительно определенная матрица весов. При этом, как и прежде, дисперсия обобщенной главной компоненты $z^{(i)}$ равна $\tilde{\lambda}_i$, а $z^{(i)}$ и $z^{(j)}$ при $i \neq j$ взаимно Ω-не коррелированы.

Заметим, кстати, что если в качестве матрицы весов выбрать матрицу

$$\Omega = \begin{pmatrix} \sigma_{11} & 0 & \ldots & 0 \\ 0 & \sigma_{22} & \ldots & 0 \\ 0 & 0 & \ldots & \sigma_{pp} \end{pmatrix},$$

то, как легко показать, обобщенные компоненты (в метрике Ω), построенные по исходным признакам $x^{(1)}, \ldots, x^{(p)}$, совпадут с обычными компонентами, построенными по вспомогательным безразмерным (нормированным) признакам $x^{*(1)}, \ldots, x^{*(p)}$.

Проиллюстрируем определение главных компонент на численном примере, заимствованном из [279].

Пример 13.1. По данным измерений (в мм) длины $(\widetilde{x^{(1)}})$, ширины $(\widetilde{x^{(2)}})$ и высоты $(\widetilde{x^{(3)}})$ панциря 24 особей ($n =$

$= 24$) одного из видов черепах определена выборочная ковариационная матрица

$$\widehat{\Sigma} = \begin{pmatrix} 451{,}39 & 271{,}17 & 168{,}70 \\ 271{,}17 & 171{,}73 & 103{,}29 \\ 168{,}70 & 103{,}29 & 66{,}65 \end{pmatrix}.$$

Решая, в соответствии с (13.4), кубическое уравнение (относительно λ) вида

$$\begin{vmatrix} 451{,}39-\lambda & 271{,}17 & 168{,}70 \\ 271{,}17 & 171{,}73-\lambda & 103{,}29 \\ 168{,}70 & 103{,}29 & 66{,}65-\lambda \end{vmatrix} = 0,$$

находим $\lambda_1 = 680{,}40$, $\lambda_2 = 6{,}50$, $\lambda_3 = 2{,}86$.

Подставляя последовательно численные значения λ_1, λ_2 и λ_3 в систему (13.3) и решая эти системы относительно неизвестных $l_i = (l_{i1}, l_{i2}, l_{i3})$ $(i = 1, 2, 3)$, получаем

$$l_1' = \begin{pmatrix} 0{,}8126 \\ 0{,}4955 \\ 0{,}3068 \end{pmatrix}, \quad l_2' = \begin{pmatrix} -0{,}5454 \\ 0{,}8321 \\ 0{,}1006 \end{pmatrix}, \quad l_3' = \begin{pmatrix} -0{,}2054 \\ -0{,}2491 \\ 0{,}9465 \end{pmatrix}.$$

В качестве главных компонент получаем

$$z^{(1)} = 0{,}81x^{(1)} + 0{,}50x^{(2)} + 0{,}31x^{(3)};$$

$$z^{(2)} = -0{,}55x^{(1)} + 0{,}83x^{(2)} + 0{,}10x^{(3)};$$

$$z^{(3)} = -0{,}21x^{(1)} - 0{,}25x^{(2)} + 0{,}95x^{(3)}.$$

Здесь под $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$ подразумеваются отклонения размеров длины ($x^{(1)}$), ширины ($x^{(2)}$)и высоты ($x^{(3)}$) панциря от своих средних значений.

Вычисление относительной доли суммарной дисперсии, обусловленной одной, двумя и тремя главными компонентами, в соответствии с формулой (13.9′) дает

$$q(1) = \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = 0{,}9864;$$

$$q(2) = \frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3} = 0{,}9958;$$

$$q(3) = 1.$$

Отсюда можно сделать вывод, что почти вся информация о специфике размеров панциря данного вида черепах содер-

жится в одной лишь первой главной компоненте, которую и естественно использовать при соответствующей классификации исследуемых особей.

13.3. Экстремальные свойства главных компонент. Их интерпретация

Свойство наименьшей ошибки «автопрогноза» или наилучшей самовоспроизводимости. Можно показать [293, 283, 284], что с помощью первых p' главных компонент $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$ $(p' < p)$ исходных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ достигается наилучший прогноз этих признаков среди всех прогнозов, которые можно построить с помощью p' линейных комбинаций набора из p произвольных признаков.

Поясним и уточним сказанное. Пусть требуется заменить исходный исследуемый p-мерный вектор наблюдений X на вектор $Z = (z^{(1)}, z^{(2)}, \ldots, z^{(p')})'$ меньшей размерности p', в котором каждая из компонент являлась бы линейной комбинацией p исходных (или каких-либо других, вспомогательных) признаков, теряя при этом не слишком много информации. Информативность нового вектора Z зависит от того, в какой степени p' введенных линейных комбинаций дают возможность «реконструировать» p исходных (измеряемых на объектах) признаков. Естественно полагать, что ошибка прогноза X по Z (обозначим ее σ) будет определяться так называемой остаточной дисперсионной матрицей вектора X при вычитании из него наилучшего прогноза по Z, т. е. матрицей $\Delta = (\Delta_{ij})$, где

$$\Delta_{ij} = \mathbf{E}\left\{\left(x^{(i)} - \sum_{l=1}^{p'} b_{il} z^{(l)}\right)\left(x^{(j)} - \sum_{l=1}^{p'} b_{jl} z^{(l)}\right)\right\}.$$

Здесь $\sum_{l=1}^{p'} b_{il} z^{(l)}$ — наилучший, в смысле метода наименьших квадратов, прогноз $x^{(i)}$ по компонентам $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$, т. е. $\sigma = f(\Delta)$, где $f(\Delta)$ — некоторая функция (качества предсказания) от элементов остаточной дисперсионной матрицы Δ.

В [293] решалась задача наилучшего прогноза X только в классе p' линейных комбинаций от исходных признаков $x^{(1)}, \ldots, x^{(p)}$ и рассмотрены естественные меры ошибки прогноза, такие, как

$$f(\Delta) = \mathrm{Sp}(\Delta) = \Delta_{11} + \Delta_{22} + \ldots + \Delta_{pp} \qquad (13.11)$$

$$f(\Delta) = \|\Delta\| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}\Delta_{ij}^2}. \qquad (13.12)$$

Здесь Sp (Δ) и $||\Delta||$ — соответственно *след* и *евклидова норма* матрицы Δ. С. Р. Рао показал, что функции (13.11) и (13.12) одновременно достигают минимума тогда и только тогда, когда в качестве $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$ выбраны первые p' главных компонент вектора X, причем величина ошибки прогноза σ явным образом выражается через последние $p - p'$ собственных чисел $\lambda_{p'+1}, \ldots, \lambda_p$ исходной ковариационной матрицы $\mathbf{\Sigma}$ или через последние $p - p'$ собственных чисел $\lambda_{p'+1}, \ldots, \lambda_p$ выборочной ковариационной матрицы $\widehat{\mathbf{\Sigma}}$, построенной по наблюдениям $X_1, X_2, \ldots, X_n$. В частности,

при $f(\Delta) = \mathrm{Sp}(\Delta)$: $\sigma = \lambda_{p'+1} + \lambda_{p'+2} + \ldots + \lambda_p$;

при $f(\Delta) = \|\Delta\|$: $\sigma = \sqrt{\lambda_{p'+1}^2 + \lambda_{p'+2}^2 + \ldots + \lambda_p^2}$.

В работах [283, 284] эта схема обобщена на случай *произвольных* предсказывающих признаков $y^{(1)}, y^{(2)}, \ldots, y^{(p')}$ и более широкого класса функций $f(\Delta)$ и показано, что $\min f(\Delta)$ достигается тогда и только тогда, когда в качестве искомых предсказывающих признаков $y^{(1)}, \ldots, y^{(p')}$ берутся сами исследуемые (измеряемые) признаки $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, а в качестве p' линейных комбинаций (предикторов) $z^{(1)}, z^{(2)}, \ldots, z^{(p')}$ от них выбраны первые p' главных компонент вектора X. При этом величина ошибки прогноза σ, как и прежде, определяется лишь $p - p'$ последними собственными значениями $\lambda_{p'+1}, \lambda_{p'+2}, \ldots, \lambda_p$ исходной ковариационной матрицы $\mathbf{\Sigma}$.

В эту схему укладывается, в частности, случай $f(\Delta) = |\Delta|$, в котором $\sigma = \lambda_{p'+1} \cdot \lambda_{p'+2} \cdot \ldots \cdot \lambda_p$.

Поясним идею описания (прогноза) исходных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ с помощью меньшего, чем p, числа их линейных комбинаций на примере 13.1.

В этом примере $p = 3$. Зададимся целью снизить размерность исходного факторного пространства до единицы ($p' = 1$), т. е. описать все три признака с помощью одной линейной комбинации от них.

В соответствии с описанным выше экстремальным свойством «автопрогноза» главных компонент возьмем в качестве этой единственной линейной комбинации первую главную компоненту, т. е. переменную

$$z^{(1)} = 0{,}81x^{(1)} + 0{,}50x^{(2)} + 0{,}31x^{(3)}.$$

Метод наименьших квадратов приводит к следующему правилу вычисления неизвестных коэффициентов b_{i1} [12 с. 209]:

$$b_{i1} = \frac{\operatorname{cov}(x^{(i)}, z^{(1)})}{\mathrm{D}z^{(1)}} =$$

$$= \frac{0{,}81 \operatorname{cov}(x^{(1)}, x^{(i)}) + 0{,}50 \operatorname{cov}(x^{(2)}, x^{(i)}) + 0{,}31 \operatorname{cov}(x^{(3)}, x^{(i)})}{\mathrm{D}z^{(1)}}.$$

Подставляя в эту формулу значения $\operatorname{cov}(x^{(i)}, x^{(j)})$, взятые из ковариационной матрицы Σ примера 13.7, получаем

$$x^{(1)} = b_{11} z^{(1)} + \varepsilon^{(1)} = 0{,}805 z^{(1)} + \varepsilon^{(1)};$$
$$x^{(2)} = b_{21} z^{(1)} + \varepsilon^{(2)} = 0{,}49 z^{(1)} + \varepsilon^{(2)};$$
$$x^{(3)} = b_{31} z^{(1)} + \varepsilon^{(3)} = 0{,}310 z^{(1)} + \varepsilon^{(3)},$$

где $\varepsilon^{(i)}$ — случайные (остаточные) ошибки прогноза исходных центрированных компонент по первой главной компоненте $z^{(1)}$.

Если в качестве относительной ошибки прогноза исходного признака $x^{(i)}$ по первой главной компоненте $z^{(1)}$ рассмотреть величину $\delta_i = (\mathrm{D}\varepsilon^{(i)}/\mathrm{D}x^{(i)}) \cdot 100\ \%$, то несложные подсчеты дают $\delta_1 = 2\ \%$, $\delta_2 = 1{,}2\ \%$ и $\delta_3 = 0{,}8\ \%$.

Суммарная характеристика относительной ошибки прогноза признаков $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$ по $z^{(1)}$ (в соответствии с вышеописанным) может быть подсчитана по формуле

$$\delta_{\text{сум. отн}} = 100\,\% \cdot \frac{\mathrm{Sp}(\Delta)}{\mathrm{D}(x^{(1)} + x^{(2)} + x^{(3)})} =$$

$$= 100\,\% \cdot \frac{\lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = 1{,}36\,\%.$$

Свойства наименьшего искажения геометрической структуры исходных точек (наблюдений) при их проектировании в пространство меньшей размерности p' «натянутое» на p' первых главных компонент. Всякий переход к меньшему числу (p') новых переменных $z^{(1)}, \ldots, z^{(p')}$, осуществляемый с помощью линейного преобразования (матрицы) $\mathbf{C} = (c_{ij})$, — $i = 1, 2, \ldots, p'$, $j = 1, 2, \ldots, p$, т. е.

$$z^{(i)} = \sum_{j=1}^{p} c_{ij} x^{(j)} \quad (i = \overline{1,\ p'}),$$

или в матричной записи

$$\mathbf{Z} = \mathbf{CX}, \tag{13.13}$$

удобнее будет рассматривать теперь как проекцию исследуемых наблюдений $X_1, X_2, \ldots, X_n$ из исходного факторного пространства $\Pi^p(X)$ в некоторое подпространство меньшей размерности $\Pi^{p'}(Z)$.

Геометрическая интерпретация сформулированных выше экстремальных свойств «автопрогноза» (самовоспроизводимости) главных компонент позволяет получить следующие интересные свойства.

Свойство 1. *Сумма квадратов расстояний от исходных точек-наблюдений $X_1, X_2, \ldots, X_n$ до пространства, натянутого на первые p' главных компонент, наименьшая относительно всех других подпространств размерности p', полученных с помощью произвольного линейного преобразования исходных координат.*

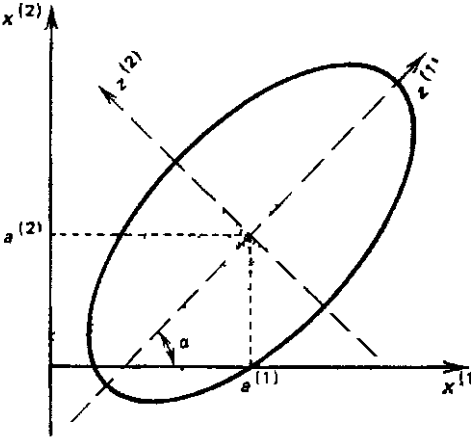


Рис. 13.2 Эллипс рассеяния исследуемых наблюдений и направление координатных осей главных компонент $z^{(1)}$ и $z^{(2)}$

Это свойство станет понятным (в свете вышеописанного экстремального свойства «автопрогноза»), если напомнить, что сумма квадратов расстояний от исходных точек до подпространства, натянутого на p' первых главных компонент, есть не что иное, как умноженная на n (общее число наблюдений) суммарная дисперсия остаточных компонент (ошибок прогноза) $\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(p)}$, следовательно, эта сумма квадратов равна $n(\lambda_{p'+1} + \lambda_{p'+2} + \ldots + \lambda_p)$. Наглядным пояснением этого свойства может служить рис. 13.2, на котором ось $z^{(1)}$ соответствует подпространству, натянутому на первую главную компоненту (т. е. $p = 2$ и $p' = 1$), а сумма квадратов расстояний до этого подпространства есть сумма перпендикуляров, опущенных из точек, изображающих наблюдения $X_i = (x_i^{(1)}, x_i^{(2)})$, на эту ось (сама ось $z^{(1)}$ может быть интерпретирована в данном случае как *линия ортогональной регрессии* $x^{(2)}$ по $x^{(1)}$, см. [7, с. 127]).

Свойство 2. *Среди всех подпространств заданной размерности p' ($p' < p$), полученных из исследуемого факторного пространства $\Pi^p(X)$ с помощью произвольного ли-*

нейного преобразования исходных координат $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, *в подпространстве, натянутом на первые* p' *главных компонент, наименее искажается сумма квадратов расстояний между всевозможными парами рассматриваемых точек-наблюдений.*

Поясним это свойство. Пусть $\Pi_C^{p'}(Z)$ — подпространство размерности p', натянутое на оси $z^{(1)}$, $z^{(2)}$, ..., $z^{(p')}$, получаемые из исходных осей $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ с помощью произвольного линейного преобразования (13.13), а Z_1, ..., Z_n — проекции исходных наблюдений X_1, ..., X_n в подпространство $\Pi_C^{p'}(Z)$, т. е. запись исходных наблюдений в координатах подпространства $\Pi_C^{p'}(Z)$. Введем в рассмотрение величины

$$M_p = \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)'(X_i - X_j);$$

$$M_{p'}(\mathbf{C}) = \sum_{i=1}^{n} \sum_{j=1}^{n} (Z_i - Z_j)'(Z_i - Z_j),$$

выражающие суммы квадратов расстояний между всевозможными парами имеющихся наблюдений соответственно в исходном пространстве $\Pi^p(X)$ и в подпространстве $\Pi_C^{p'}(Z)$.

Из простых геометрических соображений очевидно, что всегда $M_{p'}(\mathbf{C}) \leqslant M_p$ при $p' < p$.

Рассматривая в качестве меры искажения суммы квадратов попарных взаимных расстояний между точками-наблюдениями величину $M_p - M_{p'}(\mathbf{C})$, можно показать [293], что

$$M_p - M_{p'}(\mathbf{L}_{p'}) = \min_{\mathbf{C}} \{M_p - M_{p'}(\mathbf{C})\} =$$

$$= n^2 \left(\widetilde{\lambda}_{p'+1} + \widetilde{\lambda}_{p'+2} + \ldots + \widetilde{\lambda}_p\right),$$

где $\mathbf{L}(p')$ — матрица размера $p' \times p$, строками которой являются первые p' собственных векторов l_1', l_2', ..., $l_{p'}'$ исходной ковариационной матрицы $\boldsymbol{\Sigma}$ (т. е. подпространство $\Pi_{\mathbf{L}(p')}^{p'}(Z)$ является подпространством, натянутым на первые p' главных компонент вектора наблюдений X).

Свойство 3. *Среди всех подпространств заданной размерности* p' $(p' < p)$, *полученных из исследуемого факторного пространства* $\Pi^p(X)$ *с помощью произвольного линейного преобразования исходных координат* $x^{(1)}$, ..., $x^{(p)}$, *в пространстве, натянутом на первые* p' *главных компонент, наименее искажаются расстояния от рассматриваемых точек-наблюдений до их общего «центра тяжести», а также углы*

между прямыми, соединяющими всевозможные пары точек-наблюдений с их общим «центром тяжести».

Поясним это свойство. Рассмотрим матрицу G размера $(p \times n)$ «центрированных» наблюдений $x_j^{(i)} = \widetilde{x}_j^{(i)} - \overline{\widetilde{x}}^{(i)}$. Здесь, как и прежде, $\widetilde{X}_j = (\widetilde{x}_j^{(1)}, \ldots, \widetilde{x}^{(p)})$ — исходные наблюдения, а $\overline{\widetilde{x}}^{(i)} = (\widetilde{x}_1^{(i)} + \widetilde{x}_2^{(i)} + \ldots + \widetilde{x}_n^{(i)})/n$ — средняя арифметическая по всем наблюдениям i-го признака, т. е.

$$\mathbf{G} = \begin{pmatrix} x_1^{(1)} x_2^{(1)} \ldots x_n^{(1)} \\ x_1^{(2)} x_2^{(2)} \ldots x_n^{(2)} \\ \cdots\cdots\cdots \\ x_1^{(p)} x_2^{(p)} \ldots x_n^{(p)} \end{pmatrix}.$$

Введем в рассмотрение матрицу размера $(n \times n)$: $\mathbf{H} = \mathbf{G}'\mathbf{G} = (h_{jq})$, $j, q = 1, 2, \ldots, n$. Нетрудно установить геометрический смысл элементов этой матрицы:

$$h_{jj} = \sum_{i=1}^{p} \left(x_j^{(i)}\right)^2 = \sum_{i=1}^{p} \left(\widetilde{x}_j^{(i)} - \overline{\widetilde{x}}^{(i)}\right)^2 —$$

это квадрат расстояния от точки-наблюдения $\widetilde{X}_j$ до общего «центра тяжести» $\overline{\widetilde{X}}$, а

$$h_{jq} = \sum_{i=1}^{p} x_j^{(i)} x_q^{(i)} = \sum_{i=1}^{p} \left(\widetilde{x}_j^{(i)} - \overline{\widetilde{x}}^{(i)}\right)\left(\widetilde{x}_q^{(i)} - \overline{\widetilde{x}}^{(i)}\right) —$$

величина, пропорциональная косинусу угла между прямыми, соединяющими точки $\widetilde{X}_q$ и $\widetilde{X}_j$ с центром тяжести $\overline{\widetilde{X}}$.

Если рассмотреть, кроме того, матрицу $\mathbf{G}(C)$ наблюдений $Z_1, \ldots, Z_n$, являющихся проекциями исходных (центрированных) наблюдений $X_1, \ldots, X_n$ в подпространство $\Pi_C^{p'}(Z)$, и соответствующую ей матрицу $\mathbf{H}(C) = \mathbf{G}'(C) \times \mathbf{G}(C)$, то оказывается, что

$$\|\mathbf{H} - \mathbf{H}(\mathbf{L}(p'))\| = \min_{\mathbf{C}} \|\mathbf{H} - \mathbf{H}(\mathbf{C})\| =$$

$$= n^2 (\widehat{\lambda}_{p'+1}^2 + \widehat{\lambda}_{p'+2}^2 + \ldots + \lambda_p^2),$$

где под $\|\mathbf{A}\|$ понимается, как обычно, евклидова норма матрицы $\mathbf{A}$, а $\mathbf{L}(p')$ соответствует ранее введенным обозначениям.

Из описанного выше следует, что естественной мерой относительного искажения геометрической структуры исходной совокупности наблюдений при их проектировании в про-

странство меньшей размерности, натянутое на первые p' главных компонент, является величина

$$\varkappa(p') = 1 - q(p') = \frac{\lambda_{p'+1} + \ldots + \lambda_p}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

либо величина

$$\gamma(p') = \frac{\lambda_{p'+1}^2 + \ldots + \lambda_p^2}{\lambda_1^2 + \lambda_2^2 + \ldots + \lambda_p^2}.$$

При неизвестной истинной ковариационной матрице Σ ее собственные значения $\lambda_1, \ldots, \lambda_p$ следует заменить собственными значениями $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ выборочной ковариационной матрицы $\widehat{\Sigma}$ (соответственно теоретические характеристики $\varkappa$ и γ заменятся своими выборочными аналогами $\widehat{\varkappa}$ и $\widehat{\gamma}$).

13.4. Статистические свойства выборочных главных компонент; статистическая проверка некоторых гипотез

Смысл математико-статистических методов, как известно, состоит в том, чтобы по некоторой части исследуемой генеральной совокупности (по выборке или, что то же, по ограниченному ряду наблюдений $X_1, X_2, \ldots, X_n$) выносить обоснованные суждения о ее свойствах в целом.

Применительно к рассматриваемой задаче нас в первую очередь интересует, как сильно свойства и характеристики *выборочных* главных компонент могут отличаться от соответствующих свойств и характеристик главных компонент *всей генеральной совокупности* и, в частности, как эта мера отличия зависит от объема выборочной совокупности (n), по которой эти выборочные главные компоненты были построены. Так, например, для изучения природы внутренних связей между характеристиками различных статей семейного бюджета потребления и для выявления небольшого числа наиболее существенных в этом смысле показателей исследователь может обследовать какое-то количество (n) семей и по полученным результатам наблюдения $X_1, X_2, \ldots, X_n$ построить главные компоненты $\widehat{z}^{(1)}, \widehat{z}^{(2)}, \ldots, \widehat{z}^{(p')}$. Однако, увеличивая объем выборки n, т. е. добавляя к имеющимся наблюдениям результаты наблюдения по дополнительно обследованным семьям, естественно ожидать, что пересчет главных компонент *с учетом добавленных наблю-*

дений, вообще говоря, изменит (хотя, быть может, и незначительно) ранее полученные значения интересующих нас характеристик: $\widehat{\lambda}_i$, $\widehat{l}_i$ ($i = 1, 2, \ldots, p$) и т.п. В то же время существует, по-видимому, такое (столь большое) n, дальнейшее увеличение которого уже не будет практически приводить к изменению основных характеристик главных компонент (другими словами, мы вправе ожидать, что главные компоненты выборок достаточно большого объема практически совпадают с главными компонентами всей генеральной совокупности).

Выяснению некоторых вопросов, связанных с оценкой близости различных выборочных ($\widehat{z}^{(i)}$, $\widehat{l}_i$, $\widehat{\lambda}_i$) и теоретических ($z^{(i)}$, l_i, λ_i) характеристик главных компонент, и посвящен настоящий параграф. Приведенные ниже результаты исследований неизменно опираются на допущение нормальности исследуемой генеральной совокупности и взаимной независимости извлеченных из нее наблюдений. Как и прежде, под $X_1, X_2, \ldots, X_n$ будем понимать *центрированные* наблюдения, которые, строго говоря, даже при независимых исходных наблюдениях уже не будут независимыми. Однако при достаточно больших n можно пренебречь этим эффектом нарушения независимости. Таким образом, $X_i \in N(\mathbf{0}, \Sigma)$, $i = 1, 2, \ldots, n$ (как следует из предыдущего, вектор средних значений $\boldsymbol{\mu} = \mathrm{E}X$ определяет лишь точку в p-мерном пространстве, в которую переносится начало координат при переходе к главным компонентам, и с самого начала будем считать этот перенос уже осуществленным).

Вспомогательные факты, относящиеся к свойствам выборочных характеристик главных компонент [16, 279, 177, 176, 236, 235, 201]. Если все характеристические корни $\lambda_1, \lambda_2, \ldots, \lambda_p$ ковариационной матрицы Σ различны, что и имеет место в большинстве приложений анализа главных компонент, то справедливо следующее:

1) *характеристические корни* $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_p$ *и соответствующие им собственные векторы* $\widehat{l}_1, \widehat{l}_2, \ldots, \widehat{l}_p$ *выборочной ковариационной матрицы* $\widehat{\Sigma}$ *являются оценками максимального правдоподобия для соответствующих теоретических характеристик* (*соответственно* $\lambda_1, \lambda_2, \ldots, \lambda_p$ *и* $l_1, l_2, \ldots, l_p$) *и обладают всеми хорошими свойствами этих оценок* (*состоятельность, асимптотическая эффективность*). Следовательно, выборочные главные компоненты $\widehat{z}^{(i)} = \widehat{l}_i X$ ($i = \overline{1, p}$) можно интерпретировать как оценки главных компо-

нент $z^{(i)}$ всей генеральной совокупности. Если среди характеристических корней $\lambda_1, \lambda_2, \ldots, \lambda_p$ встречаются равные между собой, то оценки максимального правдоподобия для λ_i и l_i определяются иначе. Аналогичные результаты имеют место и при оценке характеристических корней и соответствующих им собственных векторов корреляционной матрицы;

2) *величины* $\sqrt{n-1}\,(\widehat{\lambda}_i - \lambda_i)$ $(i = \overline{1, p})$ *асимптотически* (*по* $n \to \infty$) *нормальны со средним значением* 0 *и с дисперсией, равной* $2\lambda_i^2$, *и независимы от других выборочных характеристических корней*;

3) *вектор* $\sqrt{n-1}\,(\widehat{l}_i - l_i)'$ $(i = \overline{1, p})$ *асимптотически* (*по* $n \to \infty$) *подчиняется многомерному нормальному распределению с вектором средних значений* **0** *и с ковариационной матрицей*

$$\lambda_i \sum_{\substack{j=1 \\ (j \neq i)}}^{p} \frac{\lambda_j}{(\lambda_j - \lambda_i)^2} \cdot l_j' l_j.$$

Этот результат имеет место для всякого λ_i, отличного от всех остальных характеристических корней, каждый из которых может иметь произвольную кратность;

4) *выборочный характеристический корень* $\widehat{\lambda}_i$ *распределен асимптотически* (*по* $n \to \infty$) *независимо от компонент соответствующего ему собственного вектора* $\widehat{l}_i$ $(i = \overline{1, p})$;

5) *ковариация между r-й компонентой выборочного собственного вектора* $\widehat{l}_i$ *и q-й компонентой выборочного собственного вектора* $\widehat{l}_j$ *равна* величине

$$-\frac{\lambda_i \lambda_j l_{ir} l_{jq}}{(n-1)(\lambda_i - \lambda_j)^2}.$$

Следующее (шестое) утверждение [20] относится к весьма специфической ситуации, характеризуемой так называемым «эффектом большой размерности», когда, несмотря на достаточно большой объем выборки n, поведение выборочных характеристик обнаруживает неожиданные особенности из-за соизмеримо (*с* n) большого значения размерности p; при этом для вывода этого факта *не требуется нормальности* исходных наблюдений,

6) *если компоненты* $x^{(i)}$ *вектора наблюдений* X *взаимно независимы и пронормированы таким образом, что* $\mathrm{E}x^{(i)} = 0$ и $\mathrm{D}x^{(i)} = 1$, *причем существуют все моменты* $\mathrm{E}\,(x^{(i)})^{\nu}$, и

если объем выборки n и размерность p одновременно достаточно велики, причем

$$\lim_{n\to\infty} \frac{p(n)}{n} = c \quad (0 < c < \infty),$$

то распределение случайно выбранного из последовательности $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_p$ характеристического корня «слабо сходится»[1] *к некоторому предельному распределению (сосредоточенному на конечном отрезке), моменты которого задаются формулой*

$$\mathbf{E}\,(\widehat{\lambda})^{\nu} = 1 + \sum_{j=1}^{\nu} c^j \frac{\nu(\nu-1)(\nu-1)\ldots(\nu-j+1)(\nu-j+1)(\nu-j)}{1\cdot 2\cdot 2\ldots j\cdot j\cdot(j+1)},$$

$(\nu = 1, 2 \ldots)$

так, что $\mathbf{E}\widehat{\lambda} = 1$, $\mathbf{E}\widehat{\lambda}^2 = 1 + c$, $\mathbf{E}\widehat{\lambda}^3 = 1 + 3c + c^2$ *и т. д.* (здесь c — некоторая постоянная величина, причем $0 < c < \infty$). Примером подобного соотношения между объемом выборки и размерностью может служить задача, описанная в [9, §2 гл. V], в которой $n = 74$, а $p = 32$ (так что $p/n = 0{,}43$).

В заключение приведем два факта, относящихся к ситуациям, в которых компоненты нормального вектора наблюдений X *взаимно независимы*:

7) *пусть* $X \in N(\mu, \Sigma)$, *где ковариационная матрица имеет диагональный вид*, т. е. $\operatorname{cov}(x^{(i)}, x^{(j)}) = 0$ *при* $i \neq j$, $i, j = 1, 2, \ldots, p$. *И пусть* $|\widehat{\mathbf{R}}| = \det(\widehat{r}_{ij})$ — *определитель выборочной корреляционной матрицы, построенной по наблюдениям* $(X_1, \ldots, X_n)$. *Тогда при достаточно больших* n $(n \to \infty)$ *статистика критерия отношения правдоподобия для проверки гипотезы о диагональном виде* Σ *может быть определена в виде* $\gamma = -\left(n - \frac{2p+11}{6}\right) \ln |\widehat{\mathbf{R}}|$, *а для ее функции распределения справедливо приближенное соотношение*

$$P\{\gamma \leqslant u\} \approx P\left\{\chi^2\left(\frac{p(p-1)}{2}\right) \leqslant u\right\}$$

при относительной ошибке, не превосходящей сотых долей процента;

---

[1] Последовательность функций $F_n(x)$, в частности последовательность функций распределения, называется слабо сходящейся (при $n \to \infty$) к функции $F(x)$, если $F_n(x)$ сходится к функции $F(x)$ на множестве ее точек непрерывности.

8) *пусть наблюдения X_j извлечены из так называемой сферической p-мерной нормальной совокупности $N(\mu, \sigma^2 I)$*, т. е. *компоненты каждого из векторов X_j взаимно независимы и имеют одинаковые дисперсии $Dx_j^{(i)}$, равные σ^2. Тогда ковариационная матрица $\Sigma = \sigma^2 I$ имеет единственный корень (кратности p), оценкой максимального правдоподобия для которого является величина*

$$\widehat{\lambda} = \frac{1}{pn} \sum_{i=1}^{p} \sum_{j=1}^{n} \left(x_j^{(i)} - \bar{x}^{(i)}\right)^2,$$

причем величина $\widehat{\lambda}/\sigma^2$ распределена по закону χ^2 ($p(n-1)$).

Статистика критерия отношения правдоподобия для проверки гипотезы о сферичности распределения исследуемого вектора наблюдений имеет вид

$$\omega = \frac{\left|n\widehat{\Sigma}\right|}{\left[\frac{1}{p}\,\mathrm{Sp}\left(n\widehat{\Sigma}\right)\right]^p}$$

и при достаточно больших n ($n \to \infty$)

$$P\left\{-\left(n-1-\frac{2p^2+p+2}{6p}\right)\ln\omega < t\right\} \approx$$
$$\approx P\left\{\chi^2\left(\frac{p(p+1)}{2}-1\right) < t\right\}$$

при относительной ошибке данного приближенного соотношения, не превосходящей сотых долей процента.

Применения свойств выборочных характеристик главных компонент. Опишем некоторые методы построения разного рода интервальных оценок для интересующих нас неизвестных характеристик главных компонент и статистической проверки гипотез, относящихся к этим характеристикам:

1) *интервальная оценка (доверительный интервал) для i-го характеристического корня λ_i*. Она получается (при больших n) с учетом асимптотической нормальности статистики $\sqrt{n-1}\,(\widehat{\lambda}_i - \lambda_i)$:

$$\frac{\widehat{\lambda}_i}{1+u_{\frac{\alpha}{2}}\sqrt{\frac{2}{n-1}}} \leqslant \lambda_i \leqslant \frac{\widehat{\lambda}_i}{1-u_{\frac{\alpha}{2}}\sqrt{\frac{2}{n-1}}}, \qquad (13.14)$$

где данное неравенство справедливо с вероятностью $1-\alpha$ (величиной α заранее задаемся), а $u_{\frac{\alpha}{2}}$ — $100\cdot\frac{\alpha}{2}\%$-ная точка стандартного нормального распределения (находится из таблиц).

Возвращаясь к примеру 13.1, по формуле (13.9) находим 95%-ный ($\alpha = 0{,}05$) доверительный интервал для наименьшего характеристического корня λ_3 по его выборочному значению $\widehat{\lambda}_3 = 2{,}86$. В этом случае $n = 24$, $u_{\frac{\alpha}{2}} = 1{,}96$, так что $1{,}81 \leqslant \lambda_3 \leqslant 6{,}78$.

Возможно обобщение асимптотического (по $n \to \infty$) доверительного интервала на случай кратных, т. е. повторяющихся, корней. Если r — кратность корня λ_i, то $100\,(1-\alpha)\%$-ный доверительный интервал для неизвестного значения λ_i задается неравенством

$$\frac{\bar{\lambda}_i}{1+u_{\frac{\alpha}{2}}\sqrt{\frac{2}{(n-1)\,r}}} \leqslant \lambda_i \leqslant \frac{\bar{\lambda}_i}{1-u_{\frac{\alpha}{2}}\sqrt{\frac{2}{(n-1)\,r}}}, \tag{13.15}$$

где

$$\bar{\lambda}_i = \frac{1}{r}\left(\widehat{\lambda}_i + \widehat{\lambda}_{i+1} + \ldots + \widehat{\lambda}_{i+r-1}\right).$$

Вопрос о том, что неизвестный характеристический корень λ_i имеет кратность и, в частности, кратность, равную r, может быть решен с помощью следующего критерия, предложенного в [176];

2) *проверка гипотезы о равенстве нескольких (а именно r) характеристических корней*: $\lambda_i = \lambda_{i+1} = \ldots = \lambda_{i+r-1}$. Очевидно, альтернативой этой гипотезе является утверждение, что не все корни среди $\lambda_i, \lambda_{i+1}, \ldots, \lambda_{i+r-1}$ равны между собой. Оказывается, в предположении справедливости проверяемой гипотезы статистика

$$\gamma_r = (n-1)\sum_{j=i}^{i+r-1}\ln\widehat{\lambda}_j + (n-1)\,r\ln\left(\frac{1}{r}\sum_{j=i}^{i+r-1}\widehat{\lambda}_j\right)$$

распределена (асимптотически по $n \to \infty$) по закону «хи-квадрат» с $r\,(r+1)/2 - 1$ степенью свободы. Поэтому гипотеза $\lambda_i = \lambda_{i+1} = \ldots = \lambda_{i+r-1}$ отвергается (с вероятностью ошибиться, равной α), если

$$\gamma_r > \chi^2_\alpha\left(\frac{r\,(r+1)}{2} - 1\right),$$

где $\chi^2_\alpha(m)$ — $100\,\alpha$ % -ная точка χ^2-распределения с m степенями свободы.

Особый интерес может представить специальный случай $i = p - r + 1$, т. е. проверка гипотезы о равенстве *последних* r собственных значений λ, что будет означать независимость и сферичность r последних признаков исследуемого вектора наблюдений.

Возвратимся к примеру 13.1. Тот факт, что оценка второго собственного значения ($\lambda_2 = 6{,}50$) попадает в доверительный интервал для λ_3 (см. выше), приводит к мысли, что, возможно, $\lambda_2 = \lambda_3$. Проверим эту гипотезу. В данном случае $n = 24$, $p = 3$, $i = 2$, $r = 2$, так что

$$\gamma_2 = -23\,(\ln 6{,}50 + \ln 2{,}86) + 46 \ln \frac{6{,}50 + 2{,}86}{2} = 3{,}70.$$

А поскольку $\chi^2_{0{,}05}(2) = 5{,}99$ и, следовательно, $\gamma_2 < \chi^2_{0{,}05}(2)$, то гипотезу $\lambda_2 = \lambda_3$ следует принять. Но тогда следует пересчитать доверительный интервал для λ_2 с учетом его кратности (в соответствии с (13.10)). Несложные подсчеты (при $\alpha = 0{,}05$ и соответственно $u_{\frac{\alpha}{2}} = u_{0{,}025} = 1{,}96$) дают: $2{,}62 \leqslant \lambda_2 \leqslant 6{,}21$, последнее неравенство будет справедливо в среднем в 95 случаях из 100;

3) *проверка гипотезы о независимости признаков* $x^{(1)}, x^{(2)}, \ldots x^{(p)}$, *являющихся компонентами вектора наблюдений* X. Такая проверка нужна для установления целесообразности применения метода главных компонент: если признаки являются взаимно независимыми, то переход к главным компонентам сведется, по существу, лишь к упорядочению исходных признаков по принципу убывания их дисперсий. Воспользуемся статистнкой критерия отношения правдоподобия для проверки гипотезы о диагональном виде ковариационной матрицы с целью проверки независимости компонент вектора наблюдений в следующем примере.

Пример 13.2. Исследовалось время, затрачиваемое работниками швейной фабрики на выполнение различных элементов операции глаженья одежды. Эту операцию можно разделить на следующие шесть элемеитов:

1) одежда размещается на гладильной доске ($x^{(1)}$);
2) разглаживаются короткие швы ($x^{(2)}$);
3) одежда перекладывается на гладильной доске ($x^{(3)}$);
4) разглаживаются длинные швы на три четверти ($x^{(4)}$);
5) разглаживаются остатки длинных швов ($x^{(5)}$);
6) одежду вешают на вешалку ($x^{(6)}$).

В этом случае X_ν представляет собой вектор измерений над ν-м индивидуумом. Компонента $x^{(i)}$ — это время, затраченное на выполнение i-го элемента операции, $n = 76$. Данные (время в секундах) обработаны, получены выборочные вектор среднего значения $\widehat{\boldsymbol{\mu}}$ и ковариационная матрица $\boldsymbol{\Sigma}$:

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} 9{,}47 \\ 25{,}56 \\ 13{,}25 \\ 31{,}44 \\ 27{,}29 \\ 8{,}70 \end{pmatrix};$$

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 2{,}57 & 0{,}85 & 1{,}56 & 1{,}79 & 1{,}33 & 0{,}42 \\ 0{,}85 & 37{,}00 & 3{,}34 & 13{,}47 & 7{,}59 & 0{,}52 \\ 1{,}56 & 3{,}34 & 8{,}44 & 5{,}77 & 2{,}00 & 0{,}50 \\ 1{,}79 & 13{,}47 & 5{,}77 & 34{,}01 & 10{,}50 & 1{,}77 \\ 1{,}33 & 7{,}59 & 2{,}00 & 10{,}50 & 23{,}01 & 3{,}43 \end{pmatrix}.$$

Выборочные стандартные отклонения равны (1,604; 6,041; 2,903; 5,832; 4,798; 2,141). Выборочная корреляционная матрица $\widehat{\mathbf{R}} = (\widehat{r_{ij}})$ имеет вид:

$$\widehat{\mathbf{R}} = \begin{pmatrix} 1{,}000 & 0{,}088 & 0{,}334 & 0{,}191 & 0{,}173 & 0{,}123 \\ 0{,}088 & 1{,}000 & 0{,}186 & 0{,}383 & 0{,}262 & 0{,}040 \\ 0{,}334 & 0{,}186 & 1{,}000 & 0{,}343 & 0{,}144 & 0{,}080 \\ 0{,}191 & 0{,}384 & 0{,}343 & 1{,}000 & 0{,}375 & 0{,}142 \\ 0{,}173 & 0{,}262 & 0{,}144 & 0{,}375 & 1{,}000 & 0{,}334 \\ 0{,}123 & 0{,}040 & 0{,}080 & 0{,}142 & 0{,}334 & 1{,}000 \end{pmatrix}.$$

Для исследователей представляет интерес проверка гипотезы о взаимной независимости шести случайных величин. Часто при изучении затрат времени предлагается новая операция, в которой элементы комбинируются иным способом. В новой операции некоторые элементы могут повторяться по нескольку раз, а некоторые могут быть выброшены. Если оказываются независимыми величины, обозначающие время, затрачиваемое на различные элементы операции, то естественно считать, что и в новой операции они останутся независимыми. Тогда распределение затрат времени на новую операцию можно будет оценить, пользуясь средними значениями и дисперсиями, вычисленными для остальных элементов. Кроме того, нас интересует возможность выделения небольшого количества вспомогательных признаков

(двух-трех), с помощью которых можно производить некоторую содержательную классификацию исследуемых работников (в том или ином смысле).

В этой задаче статистика критерия отношения правдоподобия, определенная в соответствии с п. 7 (см. с. 357), имеет вид: $\gamma = -\left(n - \frac{2p+11}{6}\right) \ln |\widehat{\mathbf{R}}| = -\frac{433}{6} \ln 0{,}472 = 54{,}1$, а $p(p-1)/2 = 15$. Задавшись уровнем значимости критерия $\alpha = 0{,}01$ (вероятность ошибочно отвергнуть проверяемую гипотезу), находим (из таблиц) величину 1%-ной точки χ^2-распределения с 15 степенями свободы: $\chi^2_{0.01}(15) = 30{,}6$. Поскольку $\gamma > \chi^2_{0.01}(15)$, то гипотезу следует отвергнуть, т. е. приходим к выводу, что значения затрат времени на различные элементы операции нельзя считать независимыми;

4) *статистическая проверка некоторых предположений (гипотез) относительно собственных векторов l_i ковариационной матрицы исследуемых признаков* ($i = 1, 2, \ldots, p$). Пусть у нас есть основания предполагать, что «нагрузки» всех признаков на первую главную компоненту равны между собой (факт симметричной зависимости первой главной компоненты от исходных признаков), т. е. $l_{11} = l_{12} = \ldots l_{1p} = \frac{1}{\sqrt{p}}$, или, напротив, что некоторые из признаков, скажем $x^{(p-1)}$ и $x^{(p)}$, вообще не влияют на первую главную компоненту (т. е. $l_{1(p-1)} = l_{1p} = 0$), в то время как остальные $p-2$ признака влияют на нее симметрично, т. е. $l_{11} = l_{12} = \ldots = l_{1(p-2)} = \frac{1}{\sqrt{p-2}}$ и т. д.

Для решения подобных вопросов можно использовать статистический критерий равенства i-го собственного вектора неизвестной ковариационной матрицы некоторому заранее заданному вектору $\tilde{l}_i$. В [176] показано, что гипотеза $l_i = \tilde{l}_i$ должна быть отвергнута (с вероятностью ошибиться, т. е. с уровнем значимости критерия, приблизительно равной α), если окажется, что

$$\gamma = (n-1)\left[\widehat{\lambda}_i \widehat{l}_i' \widehat{\Sigma}^{-1} \tilde{l}_i + \frac{1}{\widehat{\lambda}_i} \widehat{l}_i' \widehat{\Sigma} \tilde{l}_i - 2\right] > \chi^2_\alpha(p-1),$$

где подразумевается, что характеристический корень λ_i, оценка которого $\widehat{\lambda}_i$ участвует в выражении для критической статистики, имеет кратность, равную единице, а все остальные величины соответствуют ранее введенным обозначениям;

5) *проверка гипотезы о равнокоррелированности всех p исходных признаков, т. е. гипотезы* $r_{ij} = r^0$, где r_{ij} — парный коэффициент корреляции между признаком $x^{(i)}$ и признаком $x^{(j)}$ [279]. Эта гипотеза означает, что последние $p-1$ характеристических корней корреляционной матрицы равны между собой. Кроме того, постулируемый здесь специальный вид корреляционной матрицы допускает простые явные выражения в виде решений соответствующих характеристических уравнений ($\lambda_1 = 1 + (p-1)\,r^0$, $\lambda_2 = \ldots = \lambda_p = 1 - r^0$, $z^{(1)} = (x^{(1)} + x^{(2)} + \ldots + x^{(p)})/\sqrt{p}$ и т.д.) [279, с. 224].

Оказывается, гипотезу $r_{ij} = r^0$ следует отвергнуть (с вероятностью ошибиться, приблизительно равной α), если

$$\xi = \frac{n-1}{(1-\widehat{r})^2}\left[\sum_{\substack{i,j=1\\(i<j)}}^{p}(\widehat{r}_{ij}-\widehat{r})^2 - c\sum_{i=1}^{p}(\widehat{r}_i-\widehat{r})^2\right] > \chi^2_\alpha\left(\frac{(p+1)(p-2)}{2}\right),$$

где $\widehat{r}_{ij}$ — выборочные парные коэффициенты корреляции между $x^{(i)}$ и $x^{(j)}$, подсчитанные по наблюдениям $X_1, X_2, \ldots, X_n$, а

$$\widehat{r}_i = \frac{1}{p-1}\sum_{\substack{\nu=1\\(\nu\neq i)}}^{p}\widehat{r}_{i\nu},\quad \widehat{r} = \frac{2}{p(p-1)}\sum_{\substack{i,j=1\\(i\neq j)}}^{p}\widehat{r}_{ij},$$

$$c = \frac{(p-1)^2(2-\widehat{r})\,\widehat{r}}{p-(p-2)(1-\widehat{r})^2}.$$

Возвращаясь к примеру 13.1, имеем:

$$\widehat{\mathbf{R}} = \begin{pmatrix} 1{,}0000 & 0{,}9740 & 0{,}9726 \\ 0{,}9740 & 1{,}0000 & 0{,}9655 \\ 0{,}9726 & 0{,}9655 & 1{,}0000 \end{pmatrix}.$$

Несложные подсчеты дают: $\widehat{r}_1 = 0{,}9733$, $\widehat{r}_2 = 0{,}9698$, $\widehat{r}_3 = 0{,}9691$, $\widehat{r} = 0{,}9707$, так что в конечном счете $\xi = 0{,}825$.

Задавшись уровнем значимости $\alpha = 0{,}05$ и отыскав по таблицам $\chi^2_{0{,}05}(2) = 5{,}99$, приходим к выводу, что гипотеза о равнокоррелированности всех трех исходных признаков может быть признана не противоречащей имеющимся у нас результатам наблюдения.

13.5. Главные компоненты в задачах классификации

Общие идеи использования главных компонент в задачах классификации. Дуализм в постановке задачи. Очевидно, возможность геометрической интерпретации и возможность наглядного представления исследуемых наблюдений $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})'$ $(i = 1, 2, \ldots, n)$ существенно облегчает решение задач по их классификации и, в частности, проведение таких этапов, как предварительный анализ классифицируемых наблюдений, выбор метрики, выбор начальных приближений для неизвестного числа классов k, для системы эталонных множеств E, наконец, для самого́ искомого разбиения S.

Так, например, одного взгляда на рис. 13.3, на котором изображены проекции тридцати одного ($n = 31$) восемнадцатимерного наблюдения ($p = 18$) на плоскость первых двух главных компонент (построенных по исходным 18 признакам $x^{(1)}$, $x^{(2)}$, ..., $x^{(18)}$), достаточно, чтобы обнаружить четкое распадение исследуемой совокупности наблюдений на три класса [1].

Уловить же это распадение непосредственно в исходном восемнадцатимерном пространстве $\Pi^p(X)$, очевидно, невозможно.

Источником оптимизма в отношении результатов использования такого проецирования исследуемых многомерных наблюдений на плоскость являются, как легко сообразить, геометрические экстремальные свойства главных компонент, в частности вышеупомянутые свойства 1—3, в соответствии с которыми проецирование исходной совокупности наблюдений в пространство меньшей размерности, «натянутое» на p' первых главных компонент $(p' < p)$, наименее искажает ее геометрическую конфигурацию. Однако, говоря

---

[1] Данные заимствованы из работы [19]. В ней, в частности, исследовалась возможность разбиения испытываемых экземпляров растений (томатов) в пространстве признаков, характеризующих различные процессы роста растений, на однородные группы. Эти группы должны были выявить в конечном счете наличие трудноулавливаемых различий в исходных условиях выращивания (при постановке эксперимента эти условия предполагались — и, как выяснилось, необоснованно! — одинаковыми для всех растений). При исследовании было обнаружено, что первые две главные компоненты $z^{(1)}$ и $z^{(2)}$ содержат 80 % общей суммарной дисперсии всех 18 исходных признаков. При этом первую главную компоненту ($z^{(1)}$) удалось интерпретировать как характеристику общего состояния растения, в то время как вторая главная компонента ($z^{(2)}$) характеризовала процесс фотосинтеза.

о «наименьшем искажении геометрической конфигурации» совокупности исходных данных как об одном из свойств метода главных компонент, следует предостеречь читателя от «абсолютизации» в восприятии этого тезиса. В действительности *далеко не всякие геометрические свойства исходной совокупности наилучшим образом сохраняются при проецировании в плоскость первых двух главных компонент*. Так, если при проецировании исходных данных на плоскость ста-

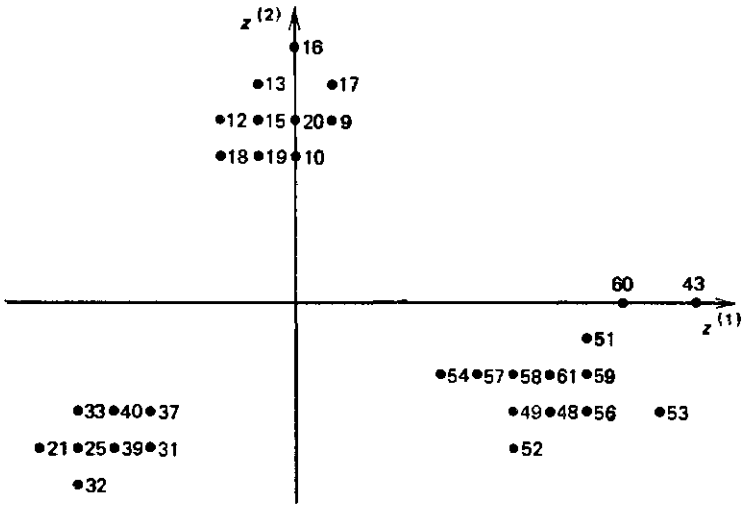


Рис. 13.3. Расположение проекций 18-мерных наблюдений на плоскость первых двух главных компонент $z^{(1)}$ и $z^{(2)}$

раются максимально сохранить разделимость существующих в исходном многомерном пространстве «сгустков», скоплений точек, то базисные оси такой плоскости будут, вообще говоря, отличаться от первых двух главных компонент. Так же, как и от осей, дающих решение аналогичной задачи при требовании (к результату проецирования) наиболее точно «выловить» резко выделяющиеся на фоне основной группы наблюдения, и т. д. Решению подобных задач, т. е. поиску плоскостей, проецирование исходных данных на которые максимально сохраняет те или иные, но наперед заданные, их геометрические свойства, посвящен раздел IV, а соответствующие методы называются *методами целенаправленного проецирования*.

Перед тем как перейти к некоторым конкретным примерам применения главных компонент в задачах классифика-

ции, обратим внимание читателя на возможную двойственность (дуализм) в интерпретации многомерного наблюдения $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})'$ вообще, и в постановке задачи при эксплуатации метода главных компонент в частности.

Действительно, если в матрице наблюдений

$$(X_1, X_2, \ldots, X_n) = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \ldots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \ldots & x_n^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ x_1^{(p)} & x_2^{(p)} & \ldots & x_n^{(p)} \end{pmatrix}$$

рассматривать в качестве наблюдения *столбцы* X_i, то классифицируемыми объектами (в количестве n штук) будут *объекты*, на каждом из которых было замерено по p признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, так или иначе характеризующих его состояние. Если же в качестве «наблюдения» рассматривать *строки* $X_\nu^{*\prime} = (x_1^{(\nu)}, x_2^{(\nu)}, \ldots, x_n^{(\nu)})$ этой матрицы, то классифицируемыми объектами будут уже сами *признаки* (в количестве p штук), рассматриваемые соответственно в n-мерном пространстве $\Pi^n(X^*)$.

Очевидно, задачи классификации в одном $\Pi^p(X)$ и в другом $\Pi^n(X^*)$ пространстве преследуют совершенно разные цели. Относительно целей классификации в пространстве $\Pi^p(X)$ выше уже говорили. Что же касается классификации в пространстве $\Pi^n(X^*)$ (классификации самих признаков), то наличие небольшого (сравнительно с p) числа однородных групп признаков позволяет сделать вывод о близости (коррелированности, взаимном дублировании) признаков, входящих в одну группу, и в конечном счете существенно снизить размерность исходного факторного пространства $\Pi^p(X)$, оставив, например, для дальнейшего рассмотрения лишь по одному представителю от каждой такой группы.

Замечание о необходимости нормировки в пространстве $\Pi^n(X)^*$. Классифицируя *признаки*, необходимо помнить, что два признака X_ν^* и X_m^* естественно считать близкими не только в случае сравнительной малости расстояния $\rho(X_\nu^*, X_m^*)$ (евклидова типа) между ними, но и в случае их достаточно тесной взаимной зависимости, например $X_\nu^* = cX_m^*$, где c – некоторый скалярный множитель. Для того чтобы это оказалось учтенным при проектировании «наблюдений» $X_1^*, X_2^*, \ldots, X_p^*$ в пространство меньшей размерности с помощью метода главных компонент, необходимо предварительно (до применения метода) соответствую-

щим образом *пронормировать* исходные данные в пространстве Π^n (X^*), например, переходя к «наблюдениям»

$$\widetilde{X}_{\nu}^{*} = \frac{X_{\nu}^{*}}{\overline{x}^{(\nu)}} \quad (\nu = 1, 2, \ldots, p),$$

где $\overline{x}^{(\nu)} = (\sum_{i=1}^{n} x_i^{(\nu)})/n$ — среднее арифметическое ν-го признака, подсчитанное по n исходным наблюдениям.

И наконец, в целях большего удобства технического представления результатов исследования (графиков, таблиц и т. п.) помимо необходимой нормировки иногда еще дополнительно *центрируют* рассматриваемые наблюдения X_{ν}^{*}, т. е. переходят в конечном счете к наблюдениям $\widetilde{\widetilde{X}}_{\nu}^{*} = \widetilde{X}_{\nu}^{*} - \widetilde{\overline{X}}_{\nu}^{*}$, где $\widetilde{\overline{X}}^{*}$ — среднее арифметическое (центр тяжести) наблюдений $\widetilde{X}_1^{*}, \widetilde{X}_2^{*}, \ldots, \widetilde{X}_p^{*}$.

В дальнейшем, как правило, будем предполагать вспомогательные операции нормировки и центрирования в пространстве Π^n (X^*) выполненными, но в целях упрощения обозначений будем опускать две верхние волнистые черточки при записи соответствующих пронормированных и процентрированных наблюдений.

Применение главных компонент при анализе структуры семейного потребления [1]. В процессе исследований решалась следующая частная задача. Объект исследований — семья. Набор измеряемых на каждом «объекте» признаков — удельные характеристики потребления (в расчете на одного члена семьи за период времени) по различным статьям расходов (табл. 13.1), всего в количестве 31 штуки ($p = 31$). На первом этапе исследований была отобрана так называемая «контрольная» выборка семей небольшого объема ($n = 106$).

Результаты проецирования 31 106-мерного наблюдения $X_{\nu}^{*'} = (x_1^{(\nu)}, x_2^{(\nu)}, \ldots, x_{106}^{(\nu)})$, $\nu = 1, 2, \ldots, 31$, на плоскость первых двух главных компонент (z_1^{*}, z_2^{*}) представлены на рис. 13.4. Если разбить исследуемые признаки на пять условных классов так, как это сделано на рисунке, то это даст пищу для достаточно естественного содержательного анализа взаимосвязей, существующих между исследуемыми признаками (лишь «расходы на кондитерские изделия» $x^{(19)}$ дали вряд ли поддающиеся содержательной интерпретации результаты проецирования: они оказались почему-то в клас-

[1] Более полно результаты этих исследований описаны в [154].

Таблица 13.1

Сумма, затрачиваемая на приобретение товара	Наименование товара (статья расхода)	Сумма, затрачиваемая на приобретение товара	Наименование товара (статья расхода)
$x^{(1)}$	Ткань	$x^{(20)}$	Общественное питание (включая расходы временно выехавших членов семьи)
$x^{(2)}$	Готовая одежда (без меховой)		
$x^{(3)}$	Меховая одежда	$x^{(21)}$	Культурно-просветительные мероприятия
$x^{(4)}$	Трикотаж	$x^{(22)}$	Транспорт
$x^{(5)}$	Обувь	$x^{(23)}$	Услуги почты и телеграфа
$x^{(6)}$	Книги, газеты	$x^{(24)}$	Жилищно-коммунальные расходы
$x^{(7)}$	Музыкальные инструменты	$x^{(25)}$	Продукты растительного происхождения
$x^{(8)}$	Спорт	$x^{(26)}$	Продукты животного происхождения
$x^{(9)}$	Мебель	$x^{(27)}$	Услуги (включая $x^{(21)}$ и $x^{(24)}$ плюс бытовые и т. п.)
$x^{(10)}$	Предметы домашнего обихода	$x^{(28)}$	Общественное питание (исключая расходы временно выехавших членов семьи)
$x^{(11)}$	Хлебобулочные изделия		
$x^{(12)}$	Овощи		
$x^{(13)}$	Мясные продукты	$x^{(29)}$	Все продовольственные товары
$x^{(14)}$	Рыбные продукты	$x^{(30)}$	Алкогольные напитки
$x^{(15)}$	Молочные продукты	$x^{(31)}$	Все промышленные товары
$x^{(16)}$	Жиры		
$x^{(17)}$	Яйца		
$x^{(18)}$	Сахар		
$x^{(19)}$	Кондитерские изделия		

се, объединяющем в себе расходы на услуги и на наиболее необходимые промышленные товары).

Применение главных компонент при анализе производительности труда рабочих. Различные показатели производительности труда $Y' = (y^{(1)}, y^{(2)}, \ldots, y^{(m)})$ характеризуют, как известно, отношение реально произведенной продукции к затратам труда на ее производство. Задача изучения зависимости показателей производительности труда от набора регулируемых (и нерегулируемых) признаков $X' =$

$= (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, характеризующих технический и организационный уровень производства, личные качества рабочих, социально-демографические условия их жизни, постоянно (и правомерно) привлекает к себе пристальное внимание исследователей. Среди различных возможных подходов к решению этой задачи выделим следующие две схемы исследования.

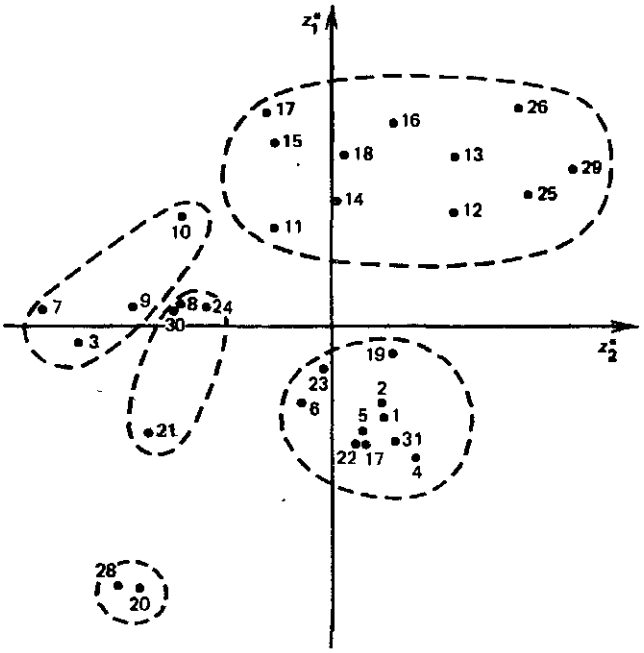


Рис. 13.4. Исследование взаимосвязей между признаками, характеризующими структуру и объем семейного потребления

Схема 1. Состоит из двух этапов:

1) разбиение исследуемой совокупности рабочих на однородные группы в пространстве объединенных признаков (X', Y'), например, с помощью главных компонент, построенных по набору признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}, y^{(1)}, \ldots, y^{(m)}$;

2) статистическое исследование зависимостей типа $Y = f_i(X)$, произведенное отдельно внутри каждой группы, выявленной на первом этапе (i — номер группы, внутри которой анализируется искомая зависимость).

Схема 2. Состоит из трех этапов:

1) разбиение исследуемой совокупности рабочих на однородные группы в пространстве признаков-аргументов $\Pi^p(X)$, например, с помощью главных компонент, построенных по набору признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$;

2) расщепление вектора признаков-аргументов $X' = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ на два подвектора: подвектор $X^{(1)'} = (x^{(1)}, x^{(2)}, \ldots, x^{(q)})$ признаков (как правило, труднорегулируемых), описывающих технический и организационный уровень производства ($q < p$), и подвектор $X^{(2)'} = (x^{(q+1)}, x^{(q+2)}, \ldots, x^{(p)})$ признаков (регулируемых), описывающих социальио-демографические условия труда. Затем разбиение исследуемой совокупности рабочих на одиородные группы $S_1^{(1)}, S_2^{(1)}, \ldots, S_{k_1}^{(1)}$ в подпространстве $\Pi^q(X^{(1)})$ «нерегулируемых» признаков, а также на однородные группы $S_1^{(2)}, S_2^{(2)}, \ldots, S_{k_2}^{(2)}$ в подпространстве $\Pi^{p-q}(X^{(2)})$ «регулируемых» признаков;

3) статистическое исследование зависимостей типа

$$Y = f_j^{(1)}(X^{(2)} \mid X^{(1)} \in S_j^{(1)}) \quad (j = 1, 2, \ldots, k_1)$$

и

$$Y = f_l^{(2)}(X^{(1)} \mid X^{(2)} \in S_l^{(2)}) \quad (l = 1, 2, \ldots, k_2),$$

произведенное отдельно внутри каждой однородной группы подпространства $\Pi^q(X^{(1)})$ (при аргументах $X^{(2)}$) и подпространства $\Pi^{p-q}(X^{(2)})$ (при аргументах $X^{(1)}$). Здесь

$$f_j^{(1)}(X^{(2)} \mid X^{(1)} \in S_j^{(1)})$$

означает *векторную* функцию от $(p-q)$ переменных $x^{(q+1)}, x^{(q+2)}, \ldots, x^{(p)}$, описывающую зависимость Y от $X^{(2)}$ при условии, что значения «нерегулируемых» аргументов $x^{(1)}, x^{(2)}, \ldots, x^{(q)}$ принадлежат области $S_j^{(1)}$. Аналогично определяется векторная функция $f_l^{(2)}$.

В [9] приводятся результаты статистического анализа исходных данных по 100 работницам-ткачихам ($n = 100$) льнокомбината «Красная текстильщица» г. Нерехта Костромской области, составляющим более 80 % всей численности ткачих комбината. Эти результаты можно рассматривать как фрагменты осуществления этапов 1 и 2 и в вышеописанных схемах исследования.

Интересный пример применения главных компонент, в прямой и двойственной постановках задачи, связанный со статистической обработкой экспертных оценок применительно к задаче классификации картин абстрактной живописи, читатель найдет в [181].

13.6. Нелинейное отображение многомерных данных в пространство низкой размерности

В некоторых случаях более точного отображения геометрической структуры исходной матрицы данных **X** в пространстве малой размерности можно добиться, используя нелинейное отображение [300, 9, 152]. Для получения таких отображений задаются тем или иным критерием (мерой) искажения $I(Z(X))$ и решают задачу на определение минимума I. Рассмотренные в данном параграфе меры искажения основаны на сравнении попарных расстояний между точками в исходном пространстве и пространстве отображения. В зависимости от выбранного критерия может получаться та или иная конфигурация точек и существенно меняется время вычисления.

13.6.1. Нелинейное отображение по критерию типа стресса. Мера искажения, рассматриваемая ниже, была предложена Сэммоном [300] и является аналогом критерия «стресса», используемого в многомерном шкалировании (см. гл. 16) [1].

$$I(Z(X)) = Q_s(Z, a) = \left(1 \Big/ \sum_{i>j}^{n-1} D_{ij}^a\right) \sum_{i>j}^{n-1} (D_{ij} - d_{ij})^2 D_{ij}^a, \tag{13.16}$$

где D_{ij} — расстояние, например, евклидово, между i-м и j-м объектами, т. е. i-й и j-й строками матрицы; d_{ij} — евклидово расстояние между образами соответствующих объектов в q-мерном пространстве.

Пусть Z_i, Z_j — q-мерные векторы координат образов объектов X_i, X_j при нелинейном отображении $X \to Z$. Расстояние d_{ij} будем считать евклидовым, т. е. $d_{ij}^2 = \sum_{k=1}^{q} (z_i^{(k)} - z_j^{(k)})^2$.

Так как евклидово расстояние не меняется при повороте осей координат, то координаты образов объектов, которые будем искать с помощью минимизации, можно считать некоррелированными (ортогональными) и центрированными $\sum_{i=1}^{n} z_i^{(k)} z_i^{(l)} = 0$, $k \neq l$. Это не меняет величины критерия, а результаты работы метода становятся более наглядными.

[1] В отличие от стресс-критериев (см. гл. 16) в критерии Сэммона значения D_{ij} не меняются в процессе работы алгоритма.

Рассмотрим сначала случай, когда $a < 0$. Тогда критерий Q_s с $a < 0$ более чувствителен к ошибкам искажения малых расстояний и менее сильно реагирует на искажение больших расстояний. Обычно рекомендуемое значение $a = -1$. При $a > 0$ лучше отображаются большие расстояния и хуже малые, так как ошибки в передаче больших расстояний сильнее влияют на значение критерия. Обычно результаты, получаемые для $a < 0$, лучше, чем для $a > 0$.

Использование двухпараметрического критерия, предложенного в [152]

$$a = \begin{cases} a_1, & \text{если } d_{ij} > D_{ij}; \\ a_2, & \text{если } d_{ij} \leqslant D_{ij}, \end{cases}$$

дает бо́льшие возможности, поскольку естественно ожидать, что можно удачно отобразить конфигурацию в пространство меньшей размерности, если искажения носят такой характер, что большие расстояния несколько увеличиваются, а малые несколько уменьшаются. Это, например, может оказаться полезным для дальнейшего использования преобразованной матрицы данных в задачах классификации, поскольку малые расстояния характерны для объектов, принадлежащих одному классу, а большие — разным. Поэтому можно ожидать, что степень разнесенности классов не слишком уменьшится в результате такого преобразования, а, возможно, и возрастет. Для получения такого эффекта надо положить $a_1 < 0$ и $a_2 > 0$.

В качестве расстояния между точками X_i и X_j в исходном пространстве признаков $x^{(1)}, \ldots, x^{(p)}$ может быть использовано любое из расстояний, перечисленных в гл. 5. Расстояние d_{ij} в пространстве образов, как уже указывалось, считается евклидовым.

Поиск образов объектов, минимизирующий значение функционала (13.16) при нелинейном отображении, осуществляется, например, с помощью итерационной градиентной процедуры:

$$z_i^{(j)}(t+1) = z_i^{(j)}(t) + b_i^{(t)} \nabla_{ij} Q_s^{(t)}, \qquad (13.17)$$

где t — номер шага итерации; $z_i^{(j)}$ — j-я координата ($j = \overline{1, q}$) образа i-го объекта ($i = \overline{1,n}$) в q-мерном пространстве; $\nabla_{ij} Q_s^{(t)}$ — первая производная Q_s по $z_i^{(j)}$; $b_i^{(t)}$ вычисляется по формуле $b_i^{(t)} = 1/2 \sum_{l=1}^{n} D_{il}^a$.

Выражение для градиента $\nabla_{ij} Q_s^{(t)}$ приведено в [11]. Пусть $Q_s^{(t)}$ — значение критерия на t-м шаге итерационной

процедуры. Остановка процедуры на t-м шаге происходит, когда выполняется хотя бы одно из условий $Q_s^{(t+1)} < \varepsilon_{\text{пор}}$, $(Q_s^{(t)} - Q_s^{(t+1)})/Q_s^{(t)} < 10^{-6}$, $t > t_{\max}$, где $t_{\max}$ — максимально допустимое число итераций; $\varepsilon_{\text{пор}}$ — допустимая точность искажения конфигурации по критерию Q_s.

В качестве начального приближения $Z^{(0)}$ для итерационной процедуры могут использоваться, например, проекции объектов на главные компоненты. Размерность пространства образов q, допустимое количество итераций $t_{\max}$ и точность $\varepsilon_{\text{пор}}$ для градиентной процедуры считаются заданными. В работе [152] предлагается применять для минимизации (13.11) метод сопряженных градиентов, который может быть эффективнее, чем градиентная процедура (13.12).

13.6.2. Быстрое нелинейное отображение с помощью опорных точек. Рассмотренный в п. 13.5.1 алгоритм нелинейного отображения требует при реализации выполнения большого количества арифметических операций, на каждом шаге итерации количество умножений пропорционально qn^2. Формирование матрицы расстояний между точками в исходном пространстве перед началом работы алгоритма также требует порядка $\frac{1}{2} pn^2$ умножений. Число умножений на каждом шаге итерации в предлагаемом ниже алгоритме быстрого нелинейного отображения (БНО) пропорционально величине $(q + 1)\, n$, т. е. растет лишь линейно, в зависимости от числа используемых объектов. Результирующая матрица точек образов на выходе этого алгоритма может быть использована либо непосредственно для дальнейшего анализа, либо как начальная конфигурация для алгоритма нелинейного отображения из п. 13.5.1.

Идея алгоритма БНО. Алгоритм БНО [66] основан на том, что в q-мерном пространстве координаты точки Z можно однозначно определить набором евклидовых расстояний $D = (\tilde{d}_0(Z), \ldots, \tilde{d}_q(Z))'$ между Z и $(q+1)$-й соответствующим образом выбранными опорными точками $\dot{Z}_0, \ldots, \dot{Z}_q$.

Действительно, разности $\tilde{d}_i^2(Z) - \tilde{d}_0^2(Z) = Z'(\dot{Z}_0 - \dot{Z}_i)\, Z + ||\dot{Z}_i||^2 - ||\dot{Z}_0||^2$ лишь линейно зависят от координат вектора Z. Поэтому для определения Z можно использовать систему из q линейных уравнений $\mathbf{A}Z = B$, где i-я строка матрицы $\mathbf{A}$ есть $A_i = 2(\dot{Z}_0 - \dot{Z}_i)$, а i-я компонента вектора B равна $b_i = ||\dot{Z}_0||^2 - ||\dot{Z}_i||^2 + \tilde{d}_i^2(Z) - \tilde{d}_0^2(Z)$, т. е. выражается через известные величины [160].

Конечно, опорные точки $\dot{Z}_0, \ldots, \dot{Z}_q$ должны быть выбраны так, чтобы матрица $\mathbf{A}$ была невырождена.

Если опорные точки зафиксированы, то в качестве функционала качества отображения может быть использована следующая величина:

$$Q_q(Z)=\sum_{j=0}^{q}\sum_{i=1}^{n}\left(\widetilde{D}_j(X_i)-\widetilde{d}_j(Z_i)\right)^2, \qquad (13.18)$$

где $\widetilde{D}_j(X_i)$ — расстояние между объектом X_i и опорной точкой X_j в исходном пространстве, а $\widetilde{d}_j(Z_i)$ — аналогичные расстояния, но в пространстве образов.

Итерационная процедура уточнения координат точки Z_i в пространстве образов имеет вид:

$$Z_i^{(t+1)}=-\sum_{j=0}^{q} w_{ij}^{(t)} Z_j/(q+1)+Z_i^{(t)}\sum_{j=0}^{q}\left(w_{ij}^{(t)}+1\right)/(q+1),$$

где $w_{ij}^{(t)}=\left(\widetilde{D}_j(X_i)-\widetilde{d}_j\left(Z_i^{(t)}\right)\right)/\widetilde{d}_j\left(Z_i^{(t)}\right)$.

Выбор нулевого приближения. Начальные значения $Z_i^{(0)}$ $(i=\overline{1,n})$ получаются из решения n систем линейных уравнений $\mathbf{A}\,Z_i^{(0)}=B_i^{(0)}$,
где j-я компонента вектора $B_i^{(0)}$ выражается через расстояния и нормы в исходном пространстве, т. е.

$$b_{ij}^{(0)}=\|X_0\|^2-\|\dot{X}_j\|^2+\widetilde{D}_j^2(X_i)-\widetilde{D}_0^2(X_i),$$

$\dot{X}_j$ — j-я опорная точка в исходном пространстве.

Выбор системы опорных точек. Выбор системы опорных точек может производиться случайным способом. Обычно его повторяют несколько раз. При этом опорные точки не должны быть слишком близки друг к другу. Поэтому новая опорная точка добавляется к ранее выбранным, только если минимальное расстояние этой точки от ранее выбранных не менее $\frac{1}{2}\,d_{\text{ср}}$, где $d_{\text{ср}}$ — среднее расстояние. В качестве окончательной конфигурации выбирается та, которая минимизирует критерий (13.18).

Другой способ состоит в разбиении матрицы $\mathbf{X}$ на $(q+1)$ сгущений с помощью алгоритма Мак-Кина и в качестве опорных выбираются точки, ближайшие к центрам классов. При наличии априорных соображений точки могут задаваться исследователем.

Указанные методы можно комбинировать и выбирать отображение либо с минимальной величиной критерия, либо наиболее подходящее с визуальной точки зрения.

Пусть $\dot{X}_0, \ldots, \dot{X}_q$ — объекты из матрицы данных, выбранные в качестве опорных. Образы опорных точек $Z_0, \ldots, Z_q$ получаются с помощью следующей процедуры.

Шаг 1. Вычисляется матрица попарных скалярных произведений размера $(q+1) \times (q+1)$

$$\mathbf{V} = (v_{ij} = (\dot{X}_i \dot{X}_j);\ i, j = \overline{0, q}).$$

Шаг 2. Вычисляются собственные числа и векторы матрицы $\mathbf{V}$.

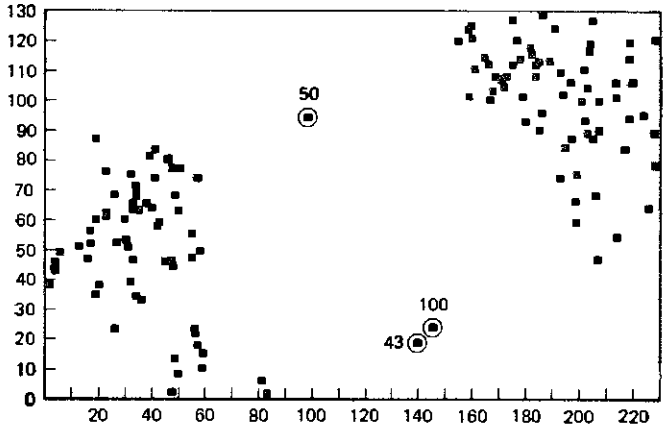


Рис 13.5 Быстрое нелинейное отображение сельскохозяйственных регионов СССР в двумерное пространство

Пусть Y_i есть i-й собственный вектор с $(q+1)$-й компонентой, а λ_i — соответствующее собственное число. Тогда i-е координаты опорных точек будут компонентами вектора $\sqrt{\lambda_i}\, Y_i$.

Пример 13.3. На рис. 13.5 предоставлены отображения, полученные с помощью алгоритма БНО. В качестве матрицы данных использовалась матрица данных из [149] ($p = 26$, $n = 130$). См. также пример 19.2. Номера опорных точек 43, 50, 100. На рисунке они обведены кружкамн. Как видим, точки разделнлнсь на трн хорошо выделенных кластера, объектнвность существования которых подтверждается их соотнесением с классификацией, предложенной в [149].

13.6.3. Быстрый алгоритм нелинейного проецирования многомерных данных. Как меру искажения геометрической конфигурации будем рассматривать здесь

$$S(\mathbf{Z}) = \frac{1}{\Sigma D_{ij}^2} \sum_{i>j}^{n} (D_{ij}^2 - d_{ij}^2)^2, \qquad (13.19)$$

где $d_{ij}^2 = \sum_{r=1}^{q} (z_i^{(r)} - z_j^{(r)})^2$.

Для дальнейшего удобно будет использовать величину

$$\tau^2(\mathbf{Z}) = \sum_{i>j}^{n} (D_{ij}^2 - d_{ij}^2)^2. \qquad (13.19')$$

Очевидно, что $S(\mathbf{Z}) = \tau^2(\mathbf{Z}) \Big/ \sum_{i>j}^{n} D_{ij}^4$.

Будем искать преобразование $Z(X) : R^p \Rightarrow R^q$, минимизирующее (13.19'), или, что то же самое, (13.19). Оказывается, для минимизации (13.19') можно построить эффективную итеративную процедуру, причем количество арифметических операций (умножений) на каждой итерации будет зависеть от n линейно, т. е. $N_{\text{умн}} = c(p, q)\, n$, где константа $c(p, q)$ от n не зависит [1].

Дальше, в силу того, что расстояния не меняются при преобразованиях переноса, не ограничивая общности, будем считать, что выполнены следующие условия: $\sum_{i=1}^{n} X_i = 0$, $\sum_{i=1}^{n} Z_i = 0$ (исходные данные и их образы центрированы).

Из аналогичных соображений (инвариантности расстояний при ортогональных преобразованиях) можно считать, что $\sum_{i=1}^{n} z_i^{(l)} z_i^{(m)} = 0$, если $l \neq m$, т. е. что переменные-образы попарно ортогональны.

Пусть дальше $s^2(z^{(k)})$ есть дисперсия (точнее, ее оценка) для $z^{(k)}$ и $s^2(x^{(k)})$ — дисперсия для $x^{(k)}$. Введем еще коэффициент сопряженности между двумя n-компонентными векторами $\mathbf{y}$ и $\mathbf{t}$

$$c(\mathbf{y}, \mathbf{t}) = \frac{1}{n} \sum_{i=1}^{n} y_i t_i.$$

[1] В [122] предложена другая процедура минимизации критерия (13.19) с числом операций, зависящим от n линейно.

Таким образом, это скалярное произведение **y** и **t**, деленное на n.

Если **y** и **t** центрированы, то c (**y**, **t**) есть просто коэффициент ковариации между переменными y и t.

Найдем теперь первую производную от τ^2 (**Z**) по $z_i^{(r)}$, т. е. $d\tau^2$ (**Z**) $/dz_i^{(r)}$, где $z_i^{(r)}$ — r-я координата образа i-го объекта ($i = \overline{1, n}$, $r = \overline{1, q}$). Имеем

$$\frac{\partial \tau^2 (\mathbf{Z})}{\partial z_i^{(r)}} = -4 \sum_{j=1}^{n} (D_{ij}^2 - d_{ij}^2)\left(z_i^{(r)} - z_j^{(r)}\right).$$

В точке глобального минимума $d\tau^2$ (**Z**)$/dz_i^{(r)} = 0$. Это приводит к следующему уравнению:

$$z_i^{(r)} \sum_{j=1}^{n} (D_{ij}^2 - d_{ij}^2) + \sum_{j=1}^{n} d_{ij}^2 z_j^{(r)} = \sum_{j=1}^{n} D_{ij}^2 z_j^{(r)}.$$

Далее имеем

$$d_{ij}^2 z_j^{(r)} = \sum_{l=1}^{q} z_j^{(r)} \left(z_i^{(l)} - z_j^{(l)}\right)^2 = \sum_{l=1}^{q} \left(z_j^{(r)} z_i^{(l)\,2} - 2 z_j^{(r)} z_i^{(l)} z_j^{(l)} + z_j^{(l)\,2} z_j^{(r)}\right).$$

Меняя порядок суммирования по l и j, получим с учетом условий ортогональности и центрированности компонент z:

$$\sum_{j=1}^{n} d_{ij}^2 z_j^{(r)} = \sum_{l=1}^{q} \sum_{j=1}^{n} z_j^{(r)} \left(z_i^{(l)} - z_j^{(l)}\right)^2 = -2ns^2 \left(z^{(r)}\right) z_i^{(r)} + n \sum_{l=1}^{q} c\left(\mathbf{z}_r^2, \mathbf{z}_r\right),$$

где векторы

$\mathbf{z}_r = \left(z_1^{(r)}, \ldots, z_n^{(r)}\right)'$;

$\mathbf{z}_r^2 = \left((z_1^{(r)})^2, \ldots, \left(z_n^{(r)}\right)^2\right)'$.

Аналогично получаем, что

$$\sum_{j=1}^{n} D_{ij}^2 z_j^{(r)} = -2n \sum_{k=1}^{p} c\left(\mathbf{x}_k \mathbf{z}_r\right) x_i^{(k)} + n \sum_{k=1}^{p} c\left(\mathbf{x}_k^2, \mathbf{z}_r\right).$$

Кроме того,

$$\sum_{j=1}^{n} D_{ij}^2 = n \| X \|^2 + n \operatorname{Sp} \mathbf{S}_X;$$

$$\sum_{j=1}^{n} d_{ij}^2 = n \| Z_i \|^2 + n \operatorname{Sp} \mathbf{S}_Z.$$

Окончательно вектор производных по компонентам z_i

$$\nabla_i \tau^2 = -4n\,(-\mathbf{B}_i Z_i + 2\mathbf{C}_{XZ} x_i + \mathbf{C}_{Z^2 Z}\, 1_q - \mathbf{C}_{X^2 Z}\, 1_p)$$

1_{q_p} — вектор размерности q (p) с единичными компонентами; $\mathbf{C}_{Z^2 Z}$ — матрица коэффициентов сопряженности $c(\mathbf{z}_l^2, \mathbf{z}_r)$ размера $q \times q$;

$$\mathbf{B}_i = 2\mathbf{V} - \mathbf{I}_q\,(\|X_i\|^2 - \|Z_i\|^2 + \mathrm{Sp}\,\mathbf{S}_X - \mathrm{Sp}\,\mathbf{S}_Z);$$

$$\mathbf{V} = \mathrm{diag}\,(s^2(z^{(1)}) \ \ldots \ s^2(z^{q}));$$

$\mathbf{C}_{X^2 Z}$ — матрица размера $q \times p$ с элементами $c_{lk} = c\,(\mathbf{x}_k^2, \mathbf{z}_l)$; $\mathbf{C}_{XZ}$ — матрица размера $q \times p$ с элементами $c_{lk} = c\,(\mathbf{x}_k \mathbf{z}_l)$.

Поскольку $x^{(k)}$ и $z^{(l)}$ центрированы, элемент c_{lk} есть не что иное, как ковариация между переменной $x^{(k)}$ и переменной $z^{(l)}$. Используем для поиска точки минимума разновидность алгоритма с переменной метрикой, для чего потребуется еще матрица вторых производных по Z_i. Непосредственным дифференцированием $\nabla_i \tau^2$ убеждаемся, что матрица вторых производных $\nabla_i^2 \tau^2$ имеет вид

$$\nabla_i^2 \tau^2 = 4n\,(\mathbf{B}_i + 2Z_i Z_i').$$

Предлагаемый алгоритм основан на применении итерационного соотношения процесса Ньютона отдельно для каждого вектора Z_i

$$Z_i^{(t+1)} = Z_i^{(t)} + \frac{1}{4n}\,(\mathbf{B}_i + 2Z_i Z_i')^{-1}\,\nabla_i \tau^2(\mathbf{Z}). \qquad (13.20)$$

Вопросы вычислительной реализации и сходимости итерационной процедуры. Дальнейшие упрощения могут быть получены за счет использования формулы Бартлетта для обращения матрицы

$$(\mathbf{B}_i + 2Z_i Z_i')^{-1} = \mathbf{B}_i^{-1} - 2\,\frac{(\mathbf{B}_i^{-1} Z_i)\,(\mathbf{B}_i^{-1} Z_i)'}{1 + 2Z_i' \mathbf{B}_i^{-1} Z_i}.$$

Матрица $\mathbf{B}_i$ диагональная, и ее обращение не представляет вычислительных трудностей. Конечно, ее диагональные элементы должны быть ненулевыми. Кроме того, для сходимости итерационного процесса (13 20) необходимо, чтобы каждая из матриц $\mathbf{B}_i$ была положительно определенной. Необходимым и достаточным условием положительной определенности всех матриц $\mathbf{B}_i$ является выполнение соотношения

$$2 \min_{1 \leqslant k \leqslant q} s^2(z^{(k)}) > \max_{1 \leqslant i \leqslant n} (\|X_i\|^2 - \|Z_i\|^2) + (\mathrm{Sp}\,\mathbf{S}_X - \mathrm{Sp}\,\mathbf{S}_Z). \qquad (13.21)$$

В частности, в точках минимума функционала (13.19′) условие (13.21) выполняется, поскольку матрица $\nabla^2\tau^2$ там неотрицательно определена и, следовательно, неотрицательно определены ее диагональные блоки $\nabla_i^2\tau$ (хотя допустимо, что (13.21) обращается в равенство).

Оценим теперь трудоемкость вычисления **градиента** $\nabla\tau^2$ и матрицы вторых производных $\nabla^2\tau^2$.

В выражение, задающее каждый из частных градиентов $\nabla_i\tau^2$, входят матрицы C_{XZ}, C_{Z^2Z}, C_{X^2Z}, *одинаковые для всех* ($i = \overline{1, n}$).

Учитывая это, получаем после некоторых преобразований, что для градиента число операций умножения пропорционально $(3p + 2q + q^2/2 + pq)\,n$.

Что касается $\nabla_i^2\tau^2$ (**Z**), то здесь необходимо дополнительно вычислить только матрицы Z_iZ_i' ($i = \overline{1, n}$), что дает в совокупности примерно $q^2n/2$ умножений. Умножение градиента на обратную матрицу (по всем $i = \overline{1, n}$ в совокупности) дает $\sim q^2n$ умножений и обращение матрицы $\nabla_i^2\tau^2$ (**Z**) с учетом формулы Бартлетта — $(q + q^2)\,n$ умножений и делений.

Окончательно количество вычислений (умножений), связанное шагом итерации по методу Ньютона, будет $N_{\text{умн}} \sim \sim(3p + 3q + 3q^2 + pq)\,n$, т. е. является линейной функцией n.

13.6.4. Сравнение нелинейного проецирования (картирования) с линейным. На первый взгляд кажется, что уменьшение суммарного искажения геометрической конфигурации данных, которое обеспечивается нелинейным проецированием, обязательно должно привести к получению большей информации о структуре данных в исходном многомерном пространстве. Ниже приводится модельный пример, который показывает, что применение нелинейного отображения может не только не улучшить, но и ухудшить передачу деталей строения многомерной конфигурации при ее отображении в пространстве более низкой размерности.

Пример 13.4. Предположим, что в двумерном пространстве переменных $x^{(1)}$, $x^{(2)}$ ($p = 2$) точки некоторой генеральной совокупности равномерно заполняют внутренность области, образованной двумя полуокружностями, расположенной так, чтобы $Ex^{(1)} = Ex^{(2)} = 0$. Присоединим к каждой из точек этой фигуры несколько переменных $x^{(3)}$, $x^{(4)}$, ..., $x^{(p)}$, равномерно распределенных на отрезке $[0, a]$ и независимых между собой и с $x^{(1)}$ и $x^{(2)}$. Так что в результате получим p-мерные объекты. В трехмерном пространстве область, заполняемая такими трехмерными точками, име-

ет вид полутороида с прямоугольным сечением (рис. 13.6). Внешний R и внутренний r радиусы тороида выберем так, что $\mathbf{D}x^{(1)} > \mathbf{D}x^{(i)}$ и $\mathbf{D}x^{(2)} > \mathbf{D}x^{(i)}$ $(i = \overline{3, p})$. Тогда проекция на две первые главные компоненты будет просто проекцией на плоскость $x^{(1)}\,0\,x^{(2)}$ (легко проверить, что матрица ковариаций будет диагональной). Эта проекция представляет собой исходную подковообразную структуру на плоскости. Точки P_1 и P_2, лежащие симметрично плоскости $x^{(1)}\,0x^{(2)}$ соответственно на верхней и нижней плоскостях тороида

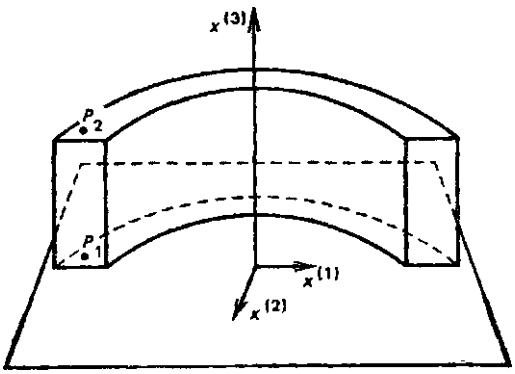


Рис. 13.6. Тороидальная структура данных в 3-мерном пространстве

для трехмерного случая, отображаются при этом в одну точку P. Однако существенная особенность исходной конфигурации — ее подковообразность — с помощью проекции на главные компоненты передается.

Процедура нелинейного проецирования стремится уменьшить искажения в передаче совокупности попарных расстояний между точками. В частности, для проекции трехмерного пространства в двумерное точки P_1 и P_2 будут разделены. Но подковообразная структура будет передана хуже.

На рис. 13.7 представлены результаты отображения моделированных данных соответственно на плоскость двух первых главных компонент и с помощью нелинейного отображения по критерию Сэммона (13.16). Использована выборка из $n = 100$ трехмерных точек $(p = 3,\ q = 2)$, равномерно распределенных внутрн тороида, представленного на рис. 13.6, с параметрами $R = 2$, $r = 1{,}8$, $a = 1{,}7$. Подковообразная структура на рис. 13.7, б существенно более «размыта», чем на рис. 13.7, а. Добавляя дополнительные «шумовые» переменные, можно добиться полного исчез-

новения подковообразной структуры при нелинейном отображенни.

Получается, что за счет улучшения передачи несущественных деталей конфигурации ухудшается отображение на-

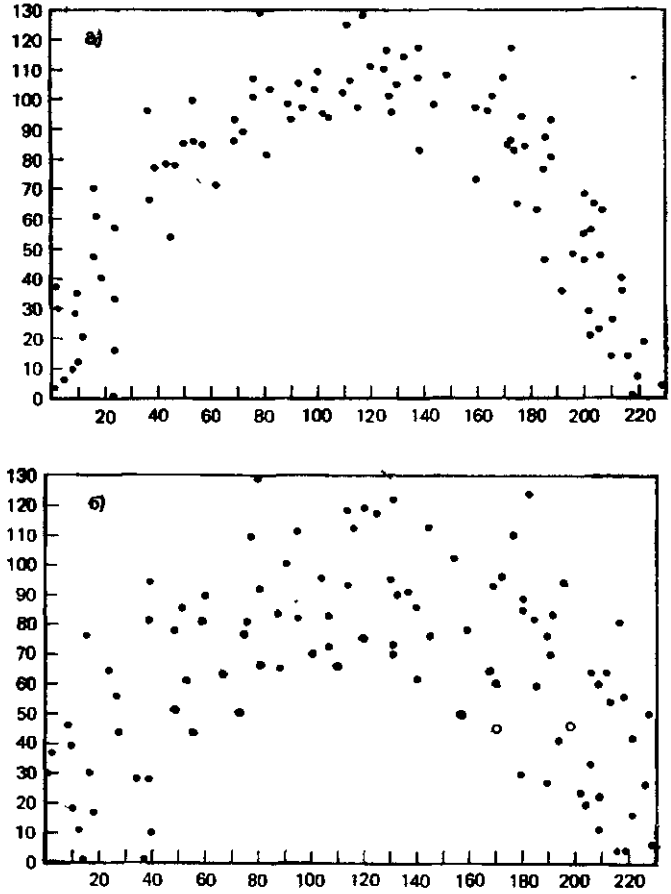


Рис. 13.7. Отображение: а) на плоскость двух первых главных компонент ($n=100$); б) нелинейное (по критерию Сэммона)

иболее интересной информации о ней. В данном случае это явление можно объяснить следующим образом. Истинное расстояние между точками X_i и X_j измеряется только с помощью координат $x^{(1)}$ и $x^{(2)}$, т. е.

$$D_{ij}^2 = (x_i^{(1)} - x_j^{(1)})^2 + (x_i^{(2)} - x_j^{(2)})^2.$$

Третья и последующие координаты $x^{(3)}$ вносят ошибку в расстояния, и в нелинейном проецировании имеем дело не с D_{ij}^2, а с расстояниями $\widehat{D}_{ij}^2 = D_{ij}^2 + \varepsilon_{ij}$, где $\varepsilon_{ij} = \sum_{k=3}^{p} (x_i^{(k)} - x_j^{(k)})^2$. При достаточно большом уровне ошибки (шума) нелинейное отображение приводит к неверной передаче особенностей исходной информации. Главные же компоненты в данной ситуации обладают лучшими фильтрационными свойствами

В то же время, если координата $x^{(3)}$ будет нести информацию о некоторой структуре данных (например, точки разделяются по $x^{(3)}$ на две хорошо обособленные группы), нелинейное отображение передает эту особенность — будем иметь две параллельные «подковы» на плоскости, а картина отображения на плоскость главных компонент не изменится.

Приведенный пример подтверждает необходимость правильного выбора переменных и метрики при использовании нелинейного проецирования и метода главных компонент, а также целесообразность использования совокупности этих методов для анализа структур данных (см. также гл. 18, 19).

ВЫВОДЫ

1. В исследовательской и практической статистической деятельности часто приходится иметь дело с исходными данными *высокой размерности*, т. е. с ситуациями, когда число регистрируемых на *каждом* из статистически обследованных объектов показателей составляет несколько десятков, а иногда — сотни и даже тысячи. В подобных ситуациях легко объяснимо желание исследователя *существенно снизить размерность анализируемого признакового пространства*, т. е. перейти от исходного набора показателей к небольшому числу *вспомогательных переменных* (которые либо отбираются из числа исходных, либо строятся по определенному правилу по совокупности исходных показателей), по которым впоследствии он мог бы достаточно точно воспроизвести интересующие его свойства анализируемого массива данных. Одним из наиболее распространенных методов снижения размерности исследуемого признакового пространства является *метод главных компонент*.

2. Имеется по меньшей мере три основных типа принципиальных предпосылок, обусловливающих возможность прак-

тически «безболезненного» перехода от большого числа исходных показателей состояния (поведения, качества, эффективности функционирования) анализируемого объекта к существенно меньшему числу наиболее информативных переменных. Это, во-первых, *дублирование информации, доставляемой сильно взаимосвязанными показателями*; во-вторых, неинформативность показателей, мало меняющихся при переходе от одного объекта к другому (*малая вариабельность показателя*); в-третьих, *возможность агрегирования*, т. е. простого или взвешенного суммирования некоторых физически однотипных показателей.

3. *Первой главной компонентой* $z^{(1)}(X)$ исследуемой системы показателей $X = (x^{(1)}, \ldots, x^{(p)})'$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая среди всех прочих нормированно-центрированных линейных комбинаций переменных $x^{(1)}, \ldots, x^{(p)}$ обладает наибольшей дисперсией. И далее: *k-й главной компонентой* ($k = 2, \ldots, p$) исследуемой системы показателей X называется такая нормированно-центрированная линейная комбинация этих показателей, которая не коррелирована с $k - 1$ предыдущими главными компонентами и среди всех прочих нормированно-центрированных и не коррелированных с предыдущими $k - 1$ главными компонентами линейных комбинаций переменных $x^{(1)}, \ldots, x^{(p)}$ обладает наибольшей дисперсией.

4. *В оптимизационной постановке задачи снижения размерности* решение, получаемое с помощью метода главных компонент, максимизирует критерий информативности, определяемый суммарной дисперсией заданного (небольшого) числа искомых вспомогательных переменных (при соответствующих условиях их нормировки). *Для вычисления k-й главной компоненты* $z^{(k)}(X)$ ($k = 1, \ldots, p$) следует найти собственный вектор $l_k = (l_{k1}, \ldots, l_{kp})$ ковариационной матрицы Σ исходного набора показателей $X = (x^{(1)}, \ldots, x^{(p)})'$, т. е. решить систему уравнений $(\Sigma - \lambda_k \mathbf{I})\, l_k' = 0$, где λ_k — k-й по величине корень (при их расположении в порядке убывания) характеристического уравнения $|\Sigma - \lambda \mathbf{I}| = 0$. Компоненты l_{kj} ($j = \overline{1, p}$) собственного вектора l_k являются искомыми весовыми коэффициентами, с помощью которых осуществляется переход от исходных показателей $x^{(1)}, \ldots x^{(p)}$ к главной компоненте $z^{(k)}(X)$, т. е. $z^{(k)}(X) = l_k \cdot X$.

5. *Основные числовые характеристики вектора* $Z = (z^{(1)}, \ldots, z^{(p)})'$ *главных компонент* могут быть выражены через основные числовые характеристики исходных показателей и собственные числа их ковариационной матрицы Σ. В частности,

$$\mathbf{E}Z = \mathbf{0};$$

$$\Sigma_Z = \mathbf{E}(Z \cdot Z') = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{pmatrix};$$

$$\sum_{j=1}^{p} \mathbf{D}z^{(j)} = \sum_{j=1}^{p} \mathbf{D}x^{(j)} = \sum_{j=1}^{p} \lambda_j; \; |\Sigma_Z| = |\Sigma|.$$

6. Вектор p' ($p' < p$) первых главных компонент $Z^{(p')}(X) = (z^{(1)}(X), \ldots, z^{(p')}(X))'$ обладает рядом *экстремальных свойств*, среди которых отметим следующие.

а) *свойство наименьшей ошибки автопрогноза или наилучшей самовоспроизводимости*: с помощью p' первых главных компонент $z^{(1)}, \ldots, z^{(p')}$ исходных показателей $x^{(1)}, \ldots, x^{(p)}$ ($p' < p$) достигается наилучший (в определенном смысле) прогноз этих показателей среди всех прогнозов, которые можно построить с помощью p' линейных комбинаций набора из p произвольных признаков,

б) *свойство наименьшего искажения некоторых геометрических характеристик совокупности исходных многомерных наблюдений* $X_1, \ldots, X_n$ при их проецировании в пространство меньшей размерности, натянутое на p' первых главных компонент $z^{(1)}, \ldots, z^{(p')}$.

7. Главные компоненты, построенные *не по истинной* ковариационной матрице Σ вектора исходных показателей $X = (x^{(1)}, \ldots x^{(p)})'$, а по ее выборочному аналогу (оценке) $\widehat{\Sigma}$, называются *выборочными главными компонентами* и в определенных (достаточно широких) условиях обладают (вместе с собственными числами и векторами матрицы $\widehat{\Sigma}$) всеми традиционными свойствами «хороших» оценок: состоятельностью, асимптотической эффективностью, асимптотической нормальностью (в условиях *растущей размерности*, т. е. в «асимптотике А. Н. Колмогорова», анализируемые выборочные характеристики могут вести себя некоторым специальным образом).

8 Геометрически определение первой главной компоненты равносильно построению новой координатной оси $0z^{(1)}$ таким образом, чтобы она шла *в направлении наибольшего разброса исходных данных*, т. е. — *в направлении вытянутости анализируемого «облака» многомерных наблюдений*. Затем среди направлений, перпендикулярных к $0z^{(1)}$, отыскивается направление «наибольшей вытянутости» $0z^{(2)}$ и т. д. Очевидно, если характер вытянутости анализируемого «об-

лака» данных в исходном признаковом пространстве существенно отличен от линейного, то *линейная* модель главных компонент может оказаться неэффективной. В подобных ситуациях исследователь должен обратиться *к нелинейным версиям метода главных компонент* (см., например, § 13.6).

9. Главные компоненты используются при решении следующих основных типов задач анализа данных:

1) *упрощение, сокращение размерностей анализируемых моделей* статистического исследования зависимостей или классификации с целью облегчения счета и интерпретации получаемых статистических выводов;

2) *наглядное представление (визуализация)* исходных многомерных данных, получаемое с помощью их проецирования в пространство, натянутое на первую, первые две или первые три главные компоненты,

3) *предварительная ортогонализация* объясняющих переменных в задачах построения регрессионных зависимостей как средство «борьбы» с мультиколлинеарностью [12, гл. 8];

4) *сжатие объемов хранимой статистической информации.*

Глава 14. МОДЕЛИ И МЕТОДЫ ФАКТОРНОГО АНАЛИЗА

14.1. Сущность модели факторного анализа, его основные задачи

Описываемые в данной главе методы основаны на общей базовой идее, в соответствии с которой структура связей между p анализируемыми признаками $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ может быть объяснена тем, что все эти переменные зависят (линейно или как-то иначе) от меньшего числа других, непосредственно не измеряемых («скрытых», «латентных») факторов $f^{(1)}$, $f^{(2)}$, ..., $f^{(p')}$ ($p' < p$), которые принято называть общими[1] и которые в большинстве моделей конструируются так, чтобы они оказались взаимно некоррелированными. При этом в общем случае, естественно, не постулируется возможность однозначного (детерминированного) восстановления значений каждого из наблюдаемых признаков $x^{(j)}$ по соот-

[1] Встречающийся еще в литературе перевод «common factor» как «простой фактор» не несет в себе главной смысловой нагрузки этого термина ведь смысл каждой из переменных $f^{(1)}$, ..., $f^{(p')}$ как раз и состоит в том, что она является *общей* для всех исходных признаков $x^{(1)}$, ..., $x^{(p)}$.

ветствующим значениям общих факторов $f^{(1)}, \ldots, f^{(p')}$ (в предположенин, что мы нх умеем вычислять): допускается, что каждый нз исходных признаков $x^{(j)}$ завнсит также и от некоторой «специфической» (для него) остаточной случайной компоненты $u^{(j)}$, которая и обусловливает статистический характер связи между $x^{(j)}$, с одной стороны, и $f^{(1)}, \ldots, f^{(p')}$ — с другой.

Конечная цель статистического исследования, проводимого с привлечением аппарата факторного анализа, как правило, состоит в выявлении и интерпретацин латентных общих факторов с одновременным противоречнвым стремлением минимизировать как их число, так и степень зависимости $x^{(j)}$ от своих специфических остаточных случайных компонент $u^{(j)}$. Как и в любой модельной схеме, эта цель может быть достигнута лишь приближенно.

В некотором смысле искомые общие факторы $f^{(1)}, \ldots, f^{(p')}$ можно считать причинами, а наблюдаемые признаки $x^{(1)}, \ldots, x^{(p)}$ — следствиями. Принято считать статистический анализ такого рода успешным, если большое число следствий удалось объяснить малым числом причин.

Таким образом, методы и модели факторного анализа нацелены на сжатие информации или, что то же, на снижение размерности исходного признакового пространства $\Pi^p(X)$. При этом из трех упомянутых в § 13.1 предпосылок возможности снижения размерности (взаимная коррелированность исходных признаков, малая «вариабельность» некоторых из них, агрегирование) методы факторного анализа базируются, в основном, на первой.

Возникновение схемы и моделей факторного анализа обязано задачам психологии, относится к началу двадцатого века и связано с именами Ч. Спирмэна, Л. Тэрстоуна, Г. Томсона [1]. Однако в силу ряда исторических причин и, в частности, из-за субъективных пристрастий и специфических научных интересов первых исследователей, работавших в данной области, вероятностно-статистические аспекты этого раздела многомерного статистического анализа долгое время оставались практически неразработанными, а интерпретации и анализу различных моделей факторного анализа была присуща некоторая неопределенность. Лишь с середины 50-х годов начинают появляться интересные результаты именно вероятиостио-статистических исследова-

[1] В качестве первой опубликованной по этой теме работы называют обычно статью Spearman C. General intelligence objectively determined and measured/Amer. J. Psychol. — 1904. — Vol. 15.— P. 201—293.

ний этого аппарата [180, 294, 185], среди которых работу Т. Андерсона и Г. Рубина следует выделить как основополагающую.

При разработке моделей факторного анализа приходится последовательно анализировать и решать следующие вопросы.

Существование модели. Далеко не для всякой заданной структуры связей между исходными признаками $X = (x^{(1)}, \ldots, x^{(p)})'$ можно (при заданном $p' < p$) построить модель факторного анализа, т. е. указать такие общие факторы $f^{(1)}, \ldots, f^{(p')}$ (или доказать их существование), которые полностью объясняли бы имеющуюся корреляцию между различными парами $x^{(i)}$ и $x^{(j)}$. При каком характере связей между исходными признаками $x^{(1)}, \ldots, x^{(p)}$, т. е. при каких корреляционных (ковариационных) матрицах $\mathbf{R} = (r_{ij})$ ($\Sigma = (\sigma_{ij})$), а также при каком соотношении между числом наблюдаемых признаков p и числом скрытых общих факторов p' ($p' < p$) сделанное допущение о наличии определенных связей между x^i ($i = 1, 2, \ldots, p$), с одной стороны, и $f^{(j)}$ ($j = 1, 2, \ldots, p'$) — с другой, является обоснованным и содержательным — в этом и заключается вопрос существования модели.

Единственность (идентификация) модели. Оказывается, что если p, Σ и p' таковы, что допускают построение модели факторного анализа, то определение соответствующих факторов $F' = (f^{(1)}, \ldots, f^{(p')})$ и коэффициентов линейного преобразования $Q = (q_{ij})$, связывающего X и F, не единственно. Спрашивается, при каких дополнительных ограничениях на матрицу преобразования $\mathbf{Q}$ и на ковариационную матрицу $V = (v_{ij})$ остаточных специфических факторов $u^{(1)}, \ldots, u^{(p)}$ определение параметров искомой модели факторного анализа будет единственным?

Алгоритмическое определение структурных параметров модели. При заданной ковариационной матрице Σ исходных признаков и известном числе общих факторов p' (и в предположении, что решение задачи определения структурных параметров Q и V существует) как конкретно вычислить неизвестные параметры модели?

Статистическое оценивание (по наблюдениям $X_1, X_2, \ldots, X_n$ и при заданном p') *неизвестных структурных параметров модели.*

Статистическая проверка ряда гипотез, связанных с природой модели и значениями ее структурных параметров, таких, как гипотеза об истинном числе p' общих факторов, гипотеза адекватности принятой модели по отношению к имеющимся результатам наблюдения, гипотеза о значимом от-

личии от нуля интересующих нас коэффициентов q_{ij} линейного преобразования и т.п.

Построение статистических оценок для ненаблюдаемых значений общих факторов $f^{(1)}, \ldots, f^{(p')}$.

14.2. Каноническая модель факторного анализа

14.2.1. Общий вид модели, ее связь с главными компонентами. Как и прежде, будем для удобства полагать исследуемые наблюдения $X_1, X_2, \ldots, X_n$ центрированными. Переход от исходных наблюдений $X_1, X_2, \ldots, X_n$ и центрированным осуществляется с помощью простого переноса начала координат в «центр тяжести» исходного множества наблюдений, т. е. $x^{(i)} = \widetilde{x}^{(i)} - \overline{\widetilde{x}}^{(i)}$, $i = 1, 2, \ldots, n$. Тогда линейная версия модели факторного анализа представляется в виде соотношений

$$\left.\begin{array}{l} X = \mathbf{Q}F + U, \\ \text{или в компонентной записи} \\ x^{(i)} = \sum\limits_{j=1}^{p'} q_{ij} f^{(j)} + u^{(i)} \quad (i = \overline{1,\ p}) \end{array}\right\}. \tag{14.1}$$

Здесь $\mathbf{Q} = (q_{ij})$ — прямоугольная матрица размера $p \times p'$ коэффициентов линейного преобразования (нагрузок общих факторов на исследуемые признаки), связывающего исследуемые признаки $x^{(i)}$ с ненаблюдаемыми (скрытыми) общими факторами $f^{(1)}, \ldots, f^{(p')}$, а вектор-столбец $U = (u^{(1)}, \ldots, u^{(p')})'$ определяет ту часть каждого из исследуемых признаков, которая не может быть объяснена общими факторами, в том числе $u^{(i)}$ включает в себя, как правило, ошибки измерения признака $x^{(i)}$.

Применительно к каждому конкретному наблюдению X_ν ($\nu = 1, 2, \ldots, n$) соотношение (14.1) дает

$$\left.\begin{array}{l} X_\nu = QF_\nu + U_\nu \\ \text{или в компонентной записи} \\ x_\nu^{(i)} = \sum\limits_{j=1}^{p'} q_{ij} f_\nu^{(j)} + u_\nu^{(i)} \quad (i = \overline{1,\ p};\ \nu = \overline{1,\ n}) \end{array}\right\}. \tag{14.1'}$$

Будем предполагать, что вектор остаточных специфических факторов U подчиняется p-мерному нормальному распределению $N(0, V)$, не зависит от F и состоит из взаимно независимых компонент, т. е. его ковариационная матрица $\mathbf{V} = \mathbf{E}(UU')$ имеет диагональный вид, где по диагонали стоят элементы $v_{ii} = \mathbf{D}u^{(i)}$.

Вектор общих факторов $F = (f^{(1)}, \ldots, f^{(p')})'$, в зависимости от содержания конкретной задачи, может интерпретироваться либо как p'-мерная нормальная случайная величина со средним $\mathbf{E}F = 0$ (в силу центрированности исходных наблюдений) и с ковариационной матрицей специального вида $\mathbf{E}(FF') = \mathbf{I}$[1], либо как вектор неизвестных неслучайных параметров, вспомогательных переменных, значения которых меняются от наблюдения к наблюдению. При последней интерпретации вектора общих факторов более правильной является запись модели в виде (14.1′), причем условия центрированности независимости и нормированности дисперсий компонент вектора F в этом случае имеют вид:

$$\frac{1}{n}\sum_{v=1}^{n} F_v = \mathbf{0}, \quad \frac{1}{n}\sum_{v=1}^{n} F_v F_v' = \mathbf{I}.$$

Однако при обоих вариантах интерпретации вектора общих факторов F исследуемый вектор наблюдений X оказывается нормально распределенной p-мерной случайной величиной: при первом варианте как линейная комбинация двух нормальных случайных векторов (F и U), а при втором варианте за счет нормальности специфических факторов $u^{(i)}$. При этом из (14.1) и из сделанных выше допущений немедленно следует, что

$$\mathbf{E}x^{(i)} = 0, \quad \begin{cases} \sigma_{ii} = \sum_{v=1}^{p'} q_{iv}^2 + v_{ii} \\ \sigma_{ij} = \sum_{v=1}^{p'} q_{iv} q_{jv} \quad (i, j = \overline{1, p}) \end{cases} \tag{14.2}$$

или в матричной записи

$$\mathbf{E}X = 0, \quad \mathbf{\Sigma} = \mathbf{QQ}' + \mathbf{V}.$$

Примером достаточно прозрачной интерпретации модели факторного анализа может служить ее формулировка в терминах так называемых интеллектуальных тестов. При этом наблюдение по признаку $x_j^{(i)}$ выражает отклонение оценки, например в баллах, данной j-му индивидууму на экзамене по i-му тесту, от некоторого среднего уровня. Естественно предположить, что в качестве ненаблюдаемых общих факторов $f^{(1)}, \ldots, f^{(p')}$, от которых будут зависеть оценки индивидуумов по всем p тестам, выступят такие факторы, как характеристика общей одаренности индивидуума $f^{(1)}$,

[1] Требование независимости компонент $f^{(i)}$ и нормированности их дисперсий объясняется в основном соображениями идентификации модели, см. § 14.1.

характеристики его математических $f^{(2)}$, технических $f^{(3)}$ или гуманитарных $f^{(4)}$ способностей.

Отметим, что соотношения (14.1) в точности воспроизводят модели множественной регрессии и дисперсионного анализа [12], в которых под $f^{(i)}$ $(i = 1, 2, \ldots, p')$ понимаются так называемые объясняющие переменные (факторы-аргументы). Однако принципиальное отличие модели факторного анализа от регрессионных схем и дисперсионного анализа состоит в том, что переменные $f^{(i)}$, выступающие в роли аргументов во всех этих моделях, не являются непосредственно наблюдаемыми в моделях факторного анализа, в то время как в регрессионном и дисперсионном анализе значения $f^{(i)}$ измеряются на статистически обследованных объектах.

Замечание 1. О связи метода главных компонент и метода факторного анализа. Рассмотрим следующую общую схему, включающую в себя в качестве частных случаев обе сравниваемые модели. Примем гипотезу, что существуют такие взаимно некоррелированные факторы $f^{(1)}, f^{(2)}, \ldots$ (быть может, в неограниченном числе), что

$$\left.\begin{aligned} x^{(1)} &= a_{11} f^{(1)} + a_{12} f^{(2)} + \ldots \\ x^{(2)} &= a_{21} f^{(1)} + a_{22} f^{(2)} + \ldots \\ &\ldots\ldots\ldots\ldots\ldots \\ x^{(p)} &= a_{p1} f^{(1)} + a_{p2} f^{(2)} + \ldots \end{aligned}\right\} \qquad (14.3)$$

или в матричной записи

$$\mathbf{X} = \mathbf{A}F,$$

где о случайных переменных $f^{(1)}$, $f^{(2)}$, ... без ограничения общности можно предположить, что $\mathbf{D} f^{(i)} = 1$.

Очевидно, представление (14.3), если оно существует, не единственно, так как переходя от F с помощью произвольного ортогонального преобразования $\mathbf{C}$ к новым переменным $Z = \mathbf{C}F$, будем иметь вместо (14.3) следующее соотношение:

$$X = \mathbf{B}Z. \qquad (14.4)$$

Исследователю не известны коэффициенты a_{ij}, но он хочет научиться наилучшим (в некотором смысле) образом аппроксимировать признаки $x^{(1)}, \ldots, x^{(p)}$ с помощью линейных функций от небольшого (заранее определенного) числа m факторов $f^{(1)}(m), \ldots, f^{(m)}(m)$, которые поэтому естественно назвать *главными* или *общими*. Аппроксимация признаков X с помощью $f^{(1)}(m), \ldots, f^{(m)}(m)$ означает представление X в виде (14.3), но с «урезанной» суммой, стоящей в правой части, т. е.

$$\widehat{X}(m) = \mathbf{A}_m F(m),$$

где $\mathbf{A}_m$ — матрица, составленная из первых m столбцов матрицы $\mathbf{A}$, а $F(m) = (f^{(1)}(m), \ldots, f^{(m)}(m))'$.

Оказывается, что, по-разному формулируя критерий оптимальности аппроксимации X с помощью $F(m)$, придем либо к главным компонентам, либо к общим факторам. Так, например, если определение элементов матрицы $\mathbf{A}_m$ подчинить идее минимизации отличия ковариационной матрицы Σ исследуемого вектора X от ковариационной матрицы $\Sigma_{\hat{X}} = \mathbf{A}_m \cdot A_m'$ аппроксимирующего вектора $\hat{X}(m)$ (в смысле минимизации евклидовой нормы $\|\Sigma - \Sigma_{\hat{X}}\|$), то $f^{(i)}(m)$ определяется пропорционально i-й главной компоненте вектора X, в частности $f^{(i)}(m) = \lambda_i^{-\frac{1}{2}} \tilde{f}^{(i)}$, где λ_i — i-й по величине характеристический корень ковариационной матрицы Σ, а $\tilde{f}^{(i)}$ — i-я главная компонента X; i-й столбец матрицы $\mathbf{A}_m$ ($i = 1, \ldots, m$) есть $\sqrt{\lambda_i} l_i$, где l_i — собственный вектор матрицы Σ, соответствующий характеристическому корню λ_i.

Если же определение аппроксимирующего вектора $\hat{X}(m) = \mathbf{B}_m F(m)$ подчинить идее максимального объяснения корреляции между исходными признаками $x^{(i)}$ и $x^{(j)}$ с помощью вспомогательных (ненаблюдаемых) факторов $f^{(1)}(m)$, $f^{(2)}(m), \ldots, f^{(m)}$ и, в частности, идее минимизации величины

$$\sum_{\substack{i,j=1 \\ (i \neq j)}}^{p} \left[\frac{\operatorname{cov}(x^{(i)}, x^{(j)}) - \operatorname{cov}\left(\hat{x}^{(i)}(m), \hat{x}^{(j)}(m)\right)}{(\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}} \right]^2 \qquad (14.5)$$

при условии неотрицательности величин $\sigma_{ii} - \mathbf{D}\hat{x}^{(i)}(m)$, то можно показать [180, 293], что i-я строка оптимальной в этом смысле матрицы преобразования $\mathbf{B}_m$ состоит из m факторных нагрузок общих факторов $f^{(1)}(m), \ldots, f^{(m)}(m)$ на i-й исходный признак $x^{(i)}$ в модели факторного анализа вида (14.1). Другими словами, сущность задачи минимизации (по $\mathbf{B}_m$ и $F(m)$) величины (14.5) состоит в следующем. Первый из m общих факторов $f^{(1)}(m)$ находится из условия, чтобы попарные корреляции между исходными признаками были как можно меньше, если влияние на них этого фактора $f^{(1)}(m)$ учтено. Следующий общий фактор $f^{(2)}(m)$ находится из условия максимального ослабления попарных корреляционных связей между исходными признаками, оставшихся после учета влияния первого общего фактора $f^{(1)}(m)$, и т. д.

Из сказанного, в частности, следует, что методы главных компонент и факторного анализа должны давать близкие результаты в тех случаях, когда главные компоненты строятся по корреляционным матрицам исходных признаков, а остаточные дисперсии v_{ii} сравнительно невелики.

Замечание 2. Вопрос о существовании модели факторного анализа. По-видимому, не всякая ковариационная матрица $\mathbf{\Sigma}$ допускает представление вида (14.2), а следовательно, не всякий вектор наблюдений X допускает интерпретацию в рамках модели факторного анализа типа (14.1). Очевидно, условия представимости вектора наблюдений X в рамках модели факторного анализа должны формулироваться в терминах свойств ковариационной матрицы $\mathbf{\Sigma}$, а также в виде некоторых соотношений между размерностью исходного пространства p и числом общих факторов p'. Одним из наиболее общих (но малоконструктивных) результатов такого рода является, например, следующее утверждение: для того чтобы вектор X допускал представление вида (14.1), необходимо и достаточно, чтобы существовала диагональная матрица $\mathbf{V}$ с неотрицательными элементами, такая, что матрица $\mathbf{\Sigma} - \mathbf{V}$ была бы неотрицательно-определенной и имела бы ранг p'. Более детальное и конструктивное исследование условий существования модели факторного анализа читатель сможет найти, например, в [180]. Заметим лишь, что изучение проблемы существования (разрешимости уравнений (14.1)) модели факторного анализа дает основу для построения различных статистических критериев адекватности модели по отношению к исследуемым наблюдениям X_1, X_2, ..., X_n.

14.2.2. Вопросы идентификации модели факторного анализа. Будем в дальнейшем предполагать, что имеется по меньшей мере одно решение ($\mathbf{Q}$, $\mathbf{V}$) уравнений (14.2). При исследовании вопроса единственности решения системы (14.2) относительно ($\mathbf{Q}$, $\mathbf{V}$) (при заданных σ_{ij}) следует различать два аспекта проблемы. Во-первых, надо понять, при каких дополнительных условиях на искомую матрицу нагрузок $\mathbf{Q}$ и на соотношение между p и p' не может существовать двух различных решений $Q^{(1)}$ и $Q^{(2)}$, таких, чтобы одно из них нельзя было бы получить из другого с помощью соответствующим образом подобранного ортогонального преобразования $\mathbf{C}$ (единственность с точностью до ортогонального преобразования или с точностью до вращения факторов). Оказывается [180], достаточным условием единственности такого рода является требование к матрице $\mathbf{Q}$, чтобы при вычеркивании из нее любой строки оставшуюся матрицу

можно было бы разделить на две подматрицы ранга p', откуда автоматически следует требование

$$p' \leqslant \frac{p-1}{2}. \tag{14.6}$$

Можно показать, что для $p' = 1$ и $p' = 2$ это условие является одновременно и необходимым, откуда, в частности, следует, что случаи $p = 2$, $p' = 1$ и $p = 4$, $p' = 2$ не допускают идентификации модели факторного анализа в указанном выше смысле (более подробное исследование идентификации этого типа можно найти в [180]).

Будем предполагать далее, что имеется по меньшей мере одно решение $(\mathbf{Q}, \mathbf{V})$ системы (14.2) и что оно единственно с точностью до ортогонального преобразования.

Вставляя в уравнения (14.2) вместо найденного решения $(\mathbf{Q}, \mathbf{V})$ другую пару матриц $(\mathbf{QC}, \mathbf{V})$, где $\mathbf{C}$ — матрица (размера $p' \times p'$) любого ортогонального преобразования, легко убедиться, что и она (эта пара матриц) удовлетворяет данной системе уравнений. Следовательно, возвращаясь к модели (14.1), получаем, что наряду с общими факторами $F = (f^{(1)}, \ldots, f^{(p')})'$ можно рассмотреть (при тех же нагрузках q_{ij}) общие факторы $Z = \mathbf{C}F$. Поскольку, как известно, ортогональное преобразование координат F геометрически означает вращение осей $f^{(1)}, \ldots, f^{(p')}$ около начала координат на некоторый угол, то получается, что при отсутствии дополнительных условий на природу искомой матрицы нагрузок $\mathbf{Q}$ общие факторы $f^{(1)}, \ldots, f^{(p')}$ могут быть определены лишь с точностью до вращения системы координат в соответствующем p'-мерном пространстве. Существует несколько вариантов дополнительных условий на класс матриц $\mathbf{Q}$, в котором следует искать решение системы (14.2), обеспечивающих уже окончательную однозначность решения $(\mathbf{Q}, \mathbf{V})$. От конкретного содержания этих условий зависит и способ численного выявления структуры искомой модели и соответственно способ статистического оценивания неизвестных параметров q_{ij}, v_{ii} и факторов $f^{(i)}$. Поэтому остановимся на них параллельно с описанием методов статистического исследования модели факторного анализа.

14.2.3. Определение структуры и статистическое исследование модели факторного анализа. Итак, в распоряжении исследователя — последовательность многомерных наблюдений $X_1, X_2, \ldots, X_n$ и с помощью модели (14.1) нужно перейти от исходных коррелированных признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, являющихся компонентами каждого из наблюдений, к меньшему числу некоррелированных вспомогательных

признаков (общих факторов) $f^{(1)}, \ldots, f^{(p')}$. Для этого надо суметь определить оценки неизвестных нагрузок $\widehat{q_{ij}}$ остаточных дисперсий $\widehat{v_{ii}}$ и, наконец, самих общих факторов $\widehat{f^{(i)}}$.

Как упоминалось, в основной модели (14.1) при $p' > 1$ оказывается слишком много неизвестных параметров для их однозначного определения. Поэтому вначале исследователь должен выбрать какую-то систему дополнительных априорных соотношений, связывающих неизвестные параметры модели, которые делают решение задачи однозначным и позволяют получить относительно простое частное решение системы (14.2). Затем он может отказаться от этих дополнительных соотношений, подбирая с помощью подходящего ортогонального преобразования (вращения осей) тот вариант оценок нагрузок $\widehat{q_{ij}}$ и остаточных дисперсий $\widehat{v_{ii}}$, который ему кажется предпочтительнее в основном в отношении возможности *содержательной интерпретации* получаемых при этом общих факторов и их нагрузок. Остановимся подробнее на основных этапах статистического исследования модели факторного анализа.

Варианты дополнительных априорных соотношений между q_{ij} и v_{ii}, постулируемых исследователем с целью однозначной идентификации анализируемой модели:

1) решение $(\mathbf{Q}, \mathbf{V})$ системы (14.2) лежит лишь в классе таких матриц $\mathbf{Q}$ и $\mathbf{V}$, для которых матрица $\mathbf{Q}'\,\mathbf{V}\,\mathbf{Q}$ имеет диагональный вид, причем диагональные элементы ее различны и упорядочены в порядке убывания[1];

2) из всех решений системы (14.2) выбирается лишь то, для которого матрица $\mathbf{Q}'\mathbf{Q}$ диагональна, причем все диагональные элементы различны и упорядочены (в порядке убывания);

3) решение системы (14.2) ищут лишь среди таких матриц $\mathbf{Q}$, которые для заранее заданной матрицы (размера $p \times p'$) $\mathbf{B} = (b_{ij}), -i = 1, \ldots, p, j = 1, \ldots, p'$ ранга p' удовлетворяют требованию

$$\mathbf{B}'\,\mathbf{Q} = \mathbf{D} = \begin{pmatrix} d_{11} & 0 & \ldots & 0 \\ d_{21} & d_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ d_{p'1} & d_{p'2} & \ldots & d_{p'p'} \end{pmatrix}.$$

[1] В некоторых случаях к этому условию добавляется требование специального вида матрицы остаточных дисперсий, а именно $\mathbf{V} = \sigma^2 \mathbf{I}$.

В частности, выбор

$$\mathbf{B}' = \begin{pmatrix} \overbrace{1 \ldots 0}^{p'} & \overbrace{0 \ldots 0}^{p-p'} \\ 1 & 0 \ldots 0 \\ 0 \ldots 1 & 0 \ldots 0 \end{pmatrix}$$

приводит к ограничению на $\mathbf{Q}$ типа

$$\mathbf{Q} = \begin{pmatrix} q_{11} & 0 & \ldots & 0 \\ q_{21} & q_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ q_{p'1} & q_{p'2} & \ldots & q_{p'p'} \\ q_{p_1} & q_{p_2} & \ldots & q_{pp'} \end{pmatrix},$$

что означает: первый исходный признак $x^{(1)}$ должен выражаться только через один первый общий фактор $f^{(1)}$, второй признак $x^{(2)}$ — через два общих фактора $f^{(1)}$ и $f^{(2)}$ и т. д.

Можно, кстати, показать, что при соответствующем выборе вспомогательной матрицы $\mathbf{B}$ определение искомых параметров модели приводит к решению ранее сформулированной задачи (14.5).

Содержательную интерпретацию условий 3) следует искать в ситуациях, когда исследователь располагает некоторой априорной информацией, из которой можно, во-первых, извлечь реальный гипотетический смысл общих факторов, и, во-вторых, постулировать наличие определенного числа нулевых элементов в матрице нагрузок $\mathbf{Q}$ (с более или менее точным указанием их «адреса»), что означает априорное отрицание зависимости исходных признаков $x^{(i)}$ от некоторых из общих факторов $f^{(j)}$ ($j = 1, 2, \ldots, p'$). Эта же идея реализуется и в других, менее формализованных вариантах дополнительных условий («простые структуры», «нулевые элементы в специфических позициях» [180]), на которых здесь не будем останавливаться.

Описание общего итерационного подхода к выявлению структуры модели факторного анализа. Конкретная реализация этого подхода зависит от выбора варианта идентифицирующих условий типа 1) — 3). Как правило, исследователю известна лишь ковариационная матрица Σ (если ее выборочное значение $\widehat{\Sigma}$, пока не будем их различать).

Логическая схема итераций следующая:

задаемся некоторым нулевым приближением $\mathbf{V}^{(0)}$ матрицы $\mathbf{V}$;

используя (14.2), получаем нулевое приближение $\Psi^{(0)} = \Sigma - V^{(0)}$ матрицы $\Psi = QQ' = \Sigma - V$;

по Ψ с помощью некоторого специального приема (см. ниже) последовательно определяем нулевые приближения $q_1^{(0)}, q_2^{(0)}, \ldots, q_{p'}^{(0)}$ для столбцов $q_1, q_2, \ldots, q_{p'}$ матрицы Q.

Затем определяем следующее (первое) приближение $V^{(1)}$ и т. д.

Специальный прием определения столбцов q_i ($i = 1, 2, \ldots, p'$) матрицы Q при известной матрице $\Psi = QQ'$ опирается на то, что матрица Ψ может быть представлена в виде $\Psi = q_1 q_1' + q_2 q_2' + \ldots + q_{p'} q_{p'}'$. Используя специфику выбранных идентифицирующих условий, определяют вначале столбец q_1. Затем переходят к матрице $\Psi_1 = \Psi - q_1 q_1' = q_2 q_2' + \ldots + q_{p'} q_{p'}'$ и определяют столбец q_2 и т.д.

Так, например, в случае условия 3) («обобщенного условия треугольности») этот прием дает

$$D = B' Q = (d_1, d_2, \ldots, d_{p'}).$$

Здесь d_i — i-й столбец матрицы D:

$$\Psi B = QQ' B = QD' = q_1 d_1' + \ldots + q_{p'} d_{p'}';$$
$$B' \Psi B = B' QQ' B = DD' = d_1 d_1' + \ldots + d_{p'} d_{p'}'.$$

Последние два матричных уравнения можно расписать в виде

$$\Psi b_i = q_1 d_{i1} + \ldots + q_i d_{ii} \quad (i = 1, 2, \ldots, p');$$
$$b_i' \Psi b_i = d_{j1} d_{i1} + \ldots + d_{jj} d_{ij} \quad (j \leqslant i,\ i = 1, \ldots, p').$$

Отсюда можно последовательно находить

$$\left.\begin{aligned} &d_{11}^2 = b_1' \Psi b_1;\ \Psi_1 = \Psi - q_1 q_1';\ q_2 = \frac{\Psi_1 b_2}{d_{22}};\\ &q_1 = \frac{\Psi b_1}{d_{11}};\ d_{22}^2 = b_2' \Psi_1 b^2;\ \Psi_2 = \Psi_1 - q_2 q_2'. \end{aligned}\right\} \qquad (14.7)$$

В случае условий идентификации типа 1) легко проверить, что столбцы $q_1, \ldots, q_{p'}$ являются первыми p' обобщенными собственными векторами (в метрике V) матрицы Ψ, т. е. являются решением уравнения

$$\Psi q_i - \lambda_i V q_i = 0 \quad (i = 1, \ldots, p'), \qquad (14.8)$$

где λ_i — i-й по величине характеристический корень уравнения

$$|\Psi - \lambda V| = 0. \qquad (14.8')$$

Поэтому общая итерационная схема определения структуры модели реализуется здесь в такой последовательности: $V^{(0)} \to \Psi^{(0)} = \Sigma - V^{(0)} \to Q^{(0)}$ — решение уравнений (14.8) $\to \Psi^{(1)} = Q^{(0)} Q^{(0)\prime} \to V^{(1)} = \Sigma - \Psi^{(1)} \to Q^{(1)}$ — решение уравнений (14.8) и т.д.

Аналогичная реализация общей итерационной схемы определения структуры модели имеет место и в случае условий идентификации типа 2) с той только разницей, что уравнения (14.8) и (14.8′) следует заменить уравнениями

$$\Psi q_i - \lambda_i q_i = 0;$$

$$|\Psi - \lambda I| = 0 \ (i = 1, 2, \ldots, p'). \qquad (14.9)$$

Статистическое оценивание факторных нагрузок q_{ij} и остаточных дисперсий v_{ii}. Оценивание производится либо методом максимального правдоподобия (см. [11, гл. 8]), либо так называемым *центроидным методом*. Первый метод используется обычно при идентифицирующих условиях типа 1) и 2). Он хотя и дает эффективные оценки для q_{ij} и v_{ii}, но требует постулирования закона распределения исследуемых величин (разработан лишь в нормальном случае), а также весьма обременительных вычислений. Центроидный метод используется при идентифицирующих условиях типа 3). Давая оценки, близкие к оценкам максимального правдоподобия, он, как и всякий непараметрический метод, является более «устойчивым» по отношению к отклонениям от нормальности исследуемых признаков и требует меньшего объема вычислений. Однако из-за определенного произвола в его процедуре, которая приведена ниже, статистическая оценка метода, исследование его выборочных свойств (в общем случае) практически невозможны. Можно представить себе проведение подобных исследований лишь в каких-то специальных случаях, один из которых намечен, например, в [180].

Общая схема реализации метода максимального правдоподобия следующая. Составляется логарифмическая функция правдоподобия как функция неизвестных параметров q_{ij} и v_{ii}, отвечающая исследуемой модели, т. е. с учетом нормальности $X_1, \ldots, X_n$ модели (14.1) и соответственно (14.2); в качестве дополнительных идентифицирующих условий берутся условия 1) или 2). С помощью дифференцирования этой функции правдоподобия по каждому из неизвестных параметров и приравнивания полученных частных производных к нулю получается система уравнений, в которой известными величинами являются выборочные ковариации

$\widehat{\sigma}_{ij}$, а также числа p и p', а неизвестными — искомые параметры q_{ij} и v_{ii}. И наконец, предлагается вычислительная (как правило, итерационная) процедура решения этой системы. Подробнее см. в [96, 161, 180]. Реализация описанной выше (для случаев 1 и 2) общей итерационной вычислительной схемы с заменой неизвестной ковариационной матрицы исходных признаков $\mathbf{\Sigma}$ ее выборочным аналогом $\widehat{\mathbf{\Sigma}}$ приведет как раз к оценкам максимального правдоподобия параметров q_{ij} и v_{ii} ($i = 1,2,\ldots, p$; $j = 1,2, \ldots, p'$). Отметим также, что в [180] при достаточно общих ограничениях доказана асимптотическая нормальность оценок максимального правдоподобия $\widehat{\mathbf{Q}}$ и $\widehat{\mathbf{V}}$, что дает основу для построения соответствующих интервальных оценок.

Как выше отмечено, центроидный метод является одним из способов реализации вычислительной схемы (14.7), приспособленной для выявления структуры модели факторного анализа и оценки неизвестных параметров в случае идентифицирующих условий типа 3). Этот метод поддается весьма простой геометрической интерпретации. Отождествим исследуемые признаки $x^{(1)}, \ldots, x^{(p)}$ с векторами, выходящими из начала координат некоторого вспомогательного p-мерного пространства, построенными таким образом, чтобы косинусы углов между $x^{(i)}$ и $x^{(j)}$ равнялись бы их парным корреляциям (r_{ij}), а длины векторов $x^{(i)}$ — стандартным отклонениям соответствующих переменных $(\sigma_{ii}^{1/2})$. Далее изменим, если необходимо, направления, т. е. знаки отдельных векторов так, чтобы как можно больше корреляций стало положительными. Тогда векторы будут иметь тенденцию к группировке в одном направлении в пучок. После этого первый общий фактор $f^{(1)}$ определяется как нормированная (т. е. как вектор единичной длины) сумма всех исходных векторов пучка, и, следовательно, он будет проходить каким-то образом через середину (центр) этого пучка; отсюда название «центроид» для общего фактора в этом случае.

Переходя затем к остаточным переменным $x^{(1)} = x^{(i)} - q_{i1}f^{(1)}$, подсчитывая ковариационную матрицу $\mathbf{\Sigma}^{(1)} = \mathbf{\Sigma} - q_1 q_1'$ для этих остаточных переменных и проделывая относительно $x^{(1)}$ и $\mathbf{\Sigma}^{(1)}$ ту же самую процедуру построения пучка и т.п., выделяем второй общий фактор («второй центроид») $f^{(2)}$ и т. д.

Формализация этих соображений приводит к следующей итерационной схеме вычислений по определению факторных нагрузок q_{ij} и остаточных дисперсий v_{ii} с учетом описанной ранее вычислительной схемы (14.7). Задаемся некоторым на-

чальным приближением $\mathbf{V}^{(0)}$ для дисперсий остатков $\mathbf{V}$. Обычно полагают [96, 161]

$$v_{ii}^{(0)} = \widehat{\sigma}_i \cdot [1 - \max_{\substack{1 \leqslant j \leqslant p \\ (j \neq i)}} |r_{ji}|].$$

Подсчитываем $\mathbf{\Psi}^{(0)} = \widehat{\mathbf{\Sigma}} - \mathbf{V}^{(0)}$. Выбираем в качестве нулевого приближения $b_1^{(0)}$ первого столбца b_1 вспомогательной матрицы $\mathbf{B}$ столбец, состоящий из одних единиц

$$b_1^{(0)} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Далее в соответствии с (14.7) определяем нулевое приближение $q_1^{(0)}$ первого столбца матрицы нагрузок

$$q_1^{(0)} = \frac{\mathbf{\Psi}^{(0)} b_1^{(0)}}{(b_1^{(0)\prime} \mathbf{\Psi}^{(0)} b_1^{(0)})^{\frac{1}{2}}}.$$

Затем вычисляется матрица $\mathbf{\Psi}_1^{(0)} = \mathbf{\Psi}^{(0)} - q_1^{(0)} q_1^{(0)\prime}$ и определяется нулевое приближение $q_2^{(0)}$ второго столбца матрицы нагрузок

$$q_2^{(0)} = \frac{\mathbf{\Psi}_1^{(0)} b_2^{(0)}}{(b_2^{(0)\prime} \mathbf{\Psi}_1^{(0)} b_2^{(0)})^{\frac{1}{2}}}, \qquad (14.10)$$

где вектор $b_2^{(0)}$ состоит только из $+1$ или -1, а знаки подбираются из условия максимизации знаменателя правой части (14.10) и т. д. Получив, таким образом, нулевое приближение $\mathbf{Q}^{(0)} = (q_1^{(0)}, \ldots, q_p^{(0)})$ для матрицы нагрузок $\mathbf{Q}$, вычисляем $\mathbf{V}^{(1)} = \mathbf{\Sigma} - \mathbf{Q}^{(0)}\mathbf{Q}^{(0)\prime}$ и переходим к следующей итерации. При этом матрица $\mathbf{B}^{(1)}$ не обязана совпадать с $\mathbf{B}^{(0)}$. Кстати, как нетрудно усмотреть из вышесказанного, i-й столбец матрицы $\mathbf{B}$ задает веса, с которыми суммируются векторы одного пучка для образования i-го общего фактора («центроида»). Поскольку смысл центроидной процедуры в простом суммировании векторов пучка, она иногда так и называется — *«процедура простого суммирования»*, то исследователю остается определить лишь нужное направление каждого из векторов пучка, т. е. знаки единиц, образующих столбцы b_i. Непосредственная ориентация (при подборе знаков у компонент вектора b_i) на максимизацию выражений $b_i^{(\nu)\prime} \mathbf{\Psi}_{i-1}^{(\nu)} b_i^{(\nu)}$ хотя и несколько сложнее реализуема, чем некоторые эвристические приемы, опирающи-

ется на анализ знаков элементов остаточных матриц Ψ_{i-1} [96, с. 41—46], но быстрее и надежнее приводит к выделению именно таких центроидов, которые при заданном p' будут обусловливать возможно большую часть общей дисперсии исходных признаков, т. е. минимизировать дисперсию остаточных компонент u_i.

Если не все исходные ковариации σ_{ij} положительны, может быть целесообразным использование и в качестве $b_1^{(0)}$ вектора, состоящего как из $+1$, так и из -1. Отметим также, что недостатком центроидного метода является зависимость центроидных нагрузок q_{ij} от шкалы, в которой измерены исходные признаки. Поэтому исходные признаки $x^{(i)}$ обычно нормируют с помощью среднеквадратических отклонений $\sigma_{ii}^{1/2}$, так что выборочная ковариационная матрица $\widehat{\Sigma}$ заменяется во всех рассуждениях выборочной корреляционной матрицей $\widehat{\mathbf{R}}$.

Анализируя описанную выше процедуру центроидного метода, нетрудно понять, что построенные таким способом общие факторы могут интерпретироваться как первые p' «условных» главных компонент матрицы $\widehat{\Sigma} - \mathbf{V}$, найденные при дополнительном условии, что компоненты соответствующих собственных векторов могут принимать лишь два значения: плюс или минус 1.

Оценка значений общих факторов. Это одна из основных задач исследования. Действительно, мало установить лишь сам факт существования небольшого числа скрыто действующих общих факторов $f^{(1)}, \ldots, f^{(p')}$, объясняющих природу взаимной коррелированности исходных признаков и основную часть их дисперсии. Желательно непосредственно определить эти общие факторы, описать их в терминах исходных признаков и постараться дать им удобную содержательную интерпретацию.

Приведем здесь идеи и результаты двух распространенных методов решения этой задачи, предложенных в разное время М. Бартлеттом (1938 г.) и Г. Томсоном (1951 г.) В обоих случаях предполагаем задачу статистического оценивания неизвестных нагрузок $\widehat{\mathbf{Q}}' = (\widehat{q}_{ij})$ и остаточных дисперсий $\widehat{\mathbf{V}} = (\widehat{v}_{ii})$ уже решенной.

Метод Бартлетта рассматривает отдельно для каждого фиксированного номера наблюдения ν ($\nu = 1, 2, \ldots, n$) модель (14.1) как регрессию признака x_ν по аргументам $\widehat{q}_{\cdot 1}$, $\widehat{q}_{\cdot 2}$, $\ldots$, $\widehat{q}_{\cdot p'}$; при этом верхний индекс $i = 1, 2, \ldots, p$ у признака (и соответствующий первый нижний индекс у на-

грузок) играет в данном случае роль номера наблюденнй в этой регрессионной схеме, так что

$$x_{\nu}^{(i)} = \sum_{j=1}^{p'} f_{\nu}^{(j)} q_{ij} + u_{\nu}^{(i)} \quad (i = 1, \ldots, p).$$

Таким образом, величины $f_{\nu}^{(1)}, f_{\nu}^{(2)}, \ldots, f_{\nu}^{(p')}$ интерпретируются как иеизвестные коэффициенты регрессии x_{ν} по $\widehat{q}_{\cdot 1}, \widehat{q}_{\cdot 2}, \ldots, \widehat{q}_{\cdot p}{}'$. В соответствии с известной техникой метода наименьших квадратов (с учетом «неравноточности» измерений, т. е. того, что, вообще говоря, $\mathrm{D}x^{(i_1)} \neq \mathrm{D}x^{(i_2)}$ при $i_1 \neq i_2$), определяющей неизвестные коэффициенты регрессии $\widehat{F}_{\nu} = (\widehat{f}_{\nu}^{(1)}, \ldots, \widehat{f}_{\nu}^{(p')})'$ из условия

$$\sum_{i=1}^{p} \frac{1}{\sigma_{ii}} \left(x_{\nu}^{(i)} - \sum_{j=1}^{p'} \widehat{f}_{\nu}^{(j)} \widehat{q}_{ij} \right)^2 = \min_{F_{\nu}} \sum_{i=1}^{p} \frac{1}{\sigma_{ii}} \times$$

$$\times \left(x_{\nu}^{(i)} - \sum_{j=1}^{p'} f_{\nu}^{(j)} \widehat{q}_{ij} \right)^2,$$

получаем

$$\widehat{F}_{\nu} = (\widehat{\mathbf{Q}}' \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{Q}})^{-1} \widehat{\mathbf{Q}}' \widehat{\mathbf{V}}^{-1} X_{\nu} \quad (\nu = 1, \ldots, n). \tag{14.11}$$

Очевидно, если исследуемый вектор наблюдений X нормален, то эти оценки являются одиовременно и оценками максимального правдоподобия. Нестрогость данного метода — в замене истинных (неизвестных нам) величин q_{ij} и v_{ii} их приближениями (оценочными) значениями $\widehat{q}_{ij}$ и $\widehat{v}_{ii}$.

Метод Томсона рассматривает модель (14.1) как бы «вывернутой наизнанку», а именно как регрессию зависимых переменных $f^{(1)}, \ldots, f^{(p')}$ по аргументам $x^{(1)}, \ldots, x^{(p)}$. Тогда коэффициенты $\widehat{c}_{ij}$ в соотношениях

$$\widehat{f}^{(i)} = \sum_{j=1}^{p} \widehat{c}_{ij} x^{(j)} \quad (i = 1, \ldots, p')$$

или в матричной записи

$$\widehat{\mathbf{F}} = \mathbf{C}X,$$

где $\mathbf{C}$ — матрица коэффициентов c_{ij} размера $p' \times p$, находят в соответствии с методом наименьших квадратов из условия

$$\sum_{\nu=1}^{n} \sum_{i=1}^{p'} \left(\widehat{f}_{\nu}^{(i)} - \sum_{j=1}^{p} \widehat{c}_{ij} x_{\nu}^{(j)} \right)^2 =$$

$$= \min_{c_{ij}} \sum_{\nu=1}^{n} \sum_{i=1}^{p'} \left(\widehat{f}_{\nu}^{(i)} - \sum_{j=1}^{p} c_{ij} x_{\nu}^{(j)} \right)^2. \qquad (14.12)$$

Поскольку решение экстремальной задачи (14.12) выписывается, как известно [16], в терминах ковариаций $x^{(i)}$ и $f^{(i)}$, то отсутствие наблюдений по зависимым переменным $f^{(i)}$ можно компенсировать знанием этих ковариаций, так как легко подсчитать, что

$$\mathbf{E}\left\{ \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(p)} \\ f^{(1)} \\ \vdots \\ f^{(p')} \end{bmatrix} (x^{(1)}, \ \ldots, \ x^{(p)}, \ f^{(1)}, \ \ldots, \ f^{(p')}) \right\} =$$

$$= \begin{pmatrix} \mathbf{QQ}' + \mathbf{V} & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{I} \end{pmatrix}.$$

Отсюда, используя известные формулы метода наименьших квадратов, получаем (с заменой матриц $\mathbf{Q}$ и $\mathbf{V}$ их выборочными аналогами)

$$F_{\nu} = (\mathbf{I} + \mathbf{\Gamma})^{-1} \widehat{\mathbf{Q}}' \ \widehat{\mathbf{V}}^{-1} X_{\nu} \ (\nu = 1, \ 2, \ \ldots, \ n), \qquad (14.13)$$

где матрица $\mathbf{\Gamma}$ (размера $p \times p$) определяется соотношением

$$\mathbf{\Gamma} = \widehat{\mathbf{Q}}' \ \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{Q}}.$$

Сравнение выражений (14.11) и (14.13) позволяет получить явное соотношение между решениями по методу Бартлетта $\widehat{F}^{(\text{Б})}$ и методу Томсона $\widehat{F}^{(\text{Т})}$.

$$\widehat{F}^{(\text{Б})} = (\mathbf{I} + \mathbf{\Gamma}^{-1}) \widehat{F}^{(\text{Т})}.$$

Если элементы матрицы $\widehat{\mathbf{Q}}' \widehat{\mathbf{V}}^{-1} \widehat{\mathbf{Q}}$ достаточно велики, то эти два метода будут давать близкие решения.

Статистическая проверка гипотез. Проверка гипотез, связанных с природой и параметрами используемой модели

факторного анализа, составляет один из необходимых моментов исследования. Теория статистнческих критериев применительно к моделям факторного анализа разработана весьма слабо. Пока удалось построить лишь так называемые критерии адекватностн модели, т. е. критерни, предназначенные для проверки гипотез типа гипотезы H_0, заключающейся в том, что исследуемый вектор наблюдений X допускает представление с помощью модели факторного анализа (14.1) с данным (заранее выбранным) числом общих факторов p'. При этом критическая статистика $\gamma(X_1, \ldots, X_n)$, т. е. функция от результатов наблюдения, по значению которой принимается решение об отклонении или непротиворечивости высказанной гипотезы H_0, зависит от вида дополнительных (идентифицирующих) условий модели. Так, если рассматривается модель с дополнительными идентифицирующими условиями вида I), т. е. дополнительно постулируется диагональность матрицы $\mathbf{\Gamma} = \widehat{\mathbf{Q}}'\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{Q}}$, то гипотеза H_0 отвергается (с вероятностью ошибиться, приблизительно равной α) в случае

$$\gamma_1(X_1, \ldots, X_n) = n(\ln|\widehat{\mathbf{V}}| + \ln|\mathbf{I} + \mathbf{\Gamma}| - \ln|\widehat{\mathbf{\Sigma}}|) > \chi^2_\alpha(\nu_1),$$

где число степеней свободы $\nu_1 = \frac{1}{2}[(p-p')^2 - (p+p')]$; его положительность обеспечивается условием (14.6), а $\chi^2_\alpha(\nu_1)$ — как и ранее, величина $100\,\alpha$ %-ной точки χ^2-распределения с ν_1 степенями свободы (находится из таблиц).

На языке ковариационных матриц гипотеза H_0 означает в данном случае, что элементы матрицы $\widehat{\mathbf{\Sigma}} - (\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}' + \widehat{\mathbf{V}})$ должны лишь статистически незначимо отличаться от нуля, или, что эквивалентно, матрица $\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{V}}$ должна иметь ранг, равный p'. А это в свою очередь означает, что последние $p - p'$ характеристических корней $\widehat{\lambda}_{p'+1}, \ldots, \widehat{\lambda}_p$ уравнения $|\widehat{\mathbf{\Sigma}} - \widehat{\mathbf{V}} - \lambda\widehat{\mathbf{V}}| = 0$ должны лишь незначимо отличаться от нуля. Статистика $\gamma_1(X_1, \ldots, X_n)$ может быть записана в терминах этих характеристических корней:

$$\gamma_1(X_1, \ldots, X_n) = n \sum_{i=p'+1}^{p} \ln(1 + \widehat{\lambda}_i).$$

Если же в качестве идентифицирующих условий дополнительно к (14.1), или, что то же, к (14.2), постулируется наличие какого-то заранее заданного числа m нулевых нагрузок q_{ij} из общего числа $p \cdot p'$ на определенных («специфи-

ческих») позициях, то гипотеза H_0 отвергается (с вероятностью ошибиться, приблизительно равной α) в случае, когда

$$\gamma_2(X_1, \ldots, X_n) = n(\ln|\widehat{\mathbf{V}}|) + \ln|\widehat{\mathbf{Q}}'\,\widehat{\mathbf{V}}^{-1}\,\widehat{\mathbf{\Sigma}}\,\mathbf{V}^{-1}\,\widehat{\mathbf{Q}}| -$$

$$- \ln|\mathbf{\Gamma}| - \ln|\widehat{\mathbf{\Sigma}}|) > \chi_\alpha^2(\nu_2),$$

где число степеней свободы $\nu_2 = \frac{1}{2}\,p\,(p-1) - (p \cdot p' - m)$.

Иногда удобнее вычислять критическую статистику $\gamma_2(X_1, \ldots, X_n)$ в терминах характеристических корней $\widehat{z}_1, \widehat{z}_2, \ldots, \widehat{z}_p$, (нумерованных в порядке убывания их величин) выборочной корреляционной матрицы $\mathbf{R}$ исследуемого вектора наблюдений X:

$$\gamma_2(X_1, \ldots, X_n) = \left(n - \frac{2p+11}{6} - \frac{2p'}{3}\right) \times$$

$$\times \left[(p-p')\,\ln\left(\frac{\sum\limits_{i=p'+1}^{p} \widehat{z}_i}{p-p'}\right) - \sum_{i=p'+1}^{p} \ln \widehat{z}_i\right].$$

Статистики $\gamma_1(X_1, \ldots, X_n)$ и $\gamma_2(X_1, \ldots, X_n)$ получены в результате реализации известной схемы критерия отношения правдоподобия.

Пользуясь этой схемой, можно построить аналогичные критерии адекватности и для некоторых специальных вариантов центроидного метода [96, с. 50]. Однако из-за слишком узких рамок такой модели эти критерии, с нашей точки зрения, не представляют достаточного интереса.

До сих пор не удалось построить многомерной решающей процедуры типа $\widehat{p'}\,(\widehat{\mathbf{\Sigma}})$, т. е. оценки для неизвестного числа общих факторов p'. В настоящее время приходится ограничиваться последовательной эксплуатацией критериев адекватности $H_0 : p' = p'_0$ (p'_0 заранее задано) при альтернативе $H_1 : p' > p'_0$. Если гипотеза H_0 отвергается, то переходят к проверке гипотезы $H'_0 : p' = p'_0 + 1$ при альтернативе $H'_1 : p' > p_0 + 1$ и т. д. Однако по уровням значимости α каждой отдельной стадии такой процедуры трудно сколько-нибудь точно судить о свойствах всей последовательной процедуры в целом.

Пользуясь асимптотической нормальностью оценок $\widehat{\mathbf{Q}}$ и $\widehat{\mathbf{V}}$, можно было бы попытаться строить критерии для проверки гипотез, касающихся значений факторных нагрузок, на-

пример, гипотез о том, что некоторые признаки не зависят от заранее определенных факторов, т. е. что на определенных местах матрицы $\widehat{\mathbf{Q}}$ стоят элементы, статистически незначимо отличающиеся от нуля. Однако построение этих критериев затруднено из-за сложности процедуры вычисления ковариационных матриц оценок $\widehat{\mathbf{Q}}$ и $\widehat{\mathbf{V}}$[1]. Правда, это затруднение можно обойти, используя в качестве приближенного решения критерий незначимого отклонения от нуля множественного коэффициента корреляции между заданным исследуемым признаком $x^{(j)}$ и заранее определенным набором факторов $\widehat{f}^{(i_1)}$, $\widehat{f}^{(i_2)}$, ..., $\widehat{f}^{(i_q)}$ $(q \leqslant p')$.

14.2.4. Факторный анализ в задачах классификации. Выше уже была отмечена близость моделей главных компонент и факторного анализа. Поэтому замечания, сформулированные в гл. 13, относящиеся к общим идеям использования главных компонент в задачах классификации и к так называемому дуализму в постановке задачи, в полной мере относятся и к модели факторного анализа.

Будет полезно пояснить это на конкретном примере с использованием специфики и терминологии именно факторного анализа.

В табл. 14.1 приведены коэффициенты корреляции между отметками по шести школьным предметам, подсчитанные по выборке из 220 учащихся [96].

В последних двух столбцах таблицы даны факторные нагрузки q_{i_1}, q_{i_2} на исследуемые признаки в бифакторной модели ($p' = 2$), подсчитанные по приведенной здесь корреляционной матрице с помощью центроидного метода. Простой анализ величин и знаков этих нагрузок склоняет нас к тому, чтобы интерпретировать первый фактор $f^{(1)}$ как фактор общей одаренности, а второй фактор $f^{(2)}$ — как фактор гуманитарной одаренности.

В прямой постановке задачи классификации (т. е. при классификации обследованных учащихся) исследователь должен был бы в первую очередь определить, как эти два общие фактора $\widehat{f}^{(1)}$ и $\widehat{f}^{(2)}$ выражаются через исходные признаки $x^{(1)}, x^{(2)}, \ldots, x^{(6)}$; затем подсчитать значения $(\widehat{f}_\nu^{(1)}, \widehat{f}_\nu^{(2)}$, $\nu = 1,2, \ldots, 220)$ этих двух факторов для каждого из обследованных учеников и, наконец, нанести 220 точек $(\widehat{f}_\nu^{(1)}$, $\widehat{f}_\nu^{(2)})$ на плоскость $\widehat{f}^{(1)}\,0\,\widehat{f}^{(2)}$. Расположение «точек-учеников»

[1] *Банников В. А.* Аппроксимация матриц и ее приложение в факторном анализе // Алгоритмическое и программное обеспечение прикладного статистического анализа. — М.: Наука, 1980.—С. 208—232.

Таблица 14.1

Содержательный смысл признака	Номер признака						Нагрузка факторов на признаки	
	1	2	3	4	5	6	q_{i1}	q_{i2}
Отметка по-гэльскому языку $x^{(1)}$	1	0,439	0,410	0,288	0,329	0,248	0,606	0,337
английскому языку $x^{(2)}$	0,439	1	0,351	0,354	0,320	0,329	0,611	0,197
истории $x^{(3)}$	0,410	0,351	1	0,164	0,190	0,181	0,458	0,384
арифметике $x^{(4)}$	0,288	0,354	0,164	1	0,595	0,570	0,683	−0,365
алгебре $x^{(5)}$	0,329	0,320	0,190	0,595	1	0,464	0,686	−0,335
геометрии $x^{(6)}$	0,248	0,329	0,181	0,470	0,464	1	0,575	−0,212

на плоскости позволило бы исследователю получить ряд вспомогательных сведений, полезных при формулировке окончательных выводов (наличие четко выраженных «сгущений точек» — классов, их число, их интерпретация и т. п.)[1]. Кстати, метод Г. Томсона (14.13) дает в качестве оценки общих факторов выражения:

$$\widehat{f}^{(1)} = 0{,}245x^{(1)} + 0{,}208x^{(2)} + 0{,}158x^{(3)} + 0{,}278x^{(4)} + $$

$$+ 0{,}271x^{(5)} + 0{,}157x^{(6)},\ \widehat{f}^{(2)} = 0{,}352x^{(1)} + 0{,}201x^{(2)} +$$

$$+ 0{,}309x^{(3)} - 0{,}351x^{(4)} - 0{,}303x^{(5)} - 0{,}126x^{(6)}.$$

При обратной (двойственной) постановке задачи, т. е. при классификации исследуемых *признаков* $x^{(1)}, x^{(2)}, \ldots, x^{(6)}$, оказывается полезной следующая геометрическая интерпретация общих факторов и исходных признаков. Рассмотрим рис. 14.1, на котором осями координат являются общие факторы $\widehat{f}^{(1)}$ и $\widehat{f}^{(2)}$, а координаты точек $(\widehat{f}_{\cdot i}^{(1)}, \widehat{f}_{\cdot i}^{(2)}) = (q_{i1}, q_{i2})$ определяются нагрузками i-го исходного признака на общие факторы ($i = 1, 2, \ldots, 6$). Соответственно точку (q_{i1}, q_{i2}) удобно интерпретировать как изображение i-го исходного признака $x^{(i)}$. Расположение точек на рис. 14.1 свидетельст-

[1] Аналогичная задача классификации ткачих при исследовании их производительности труда упоминалась в гл. 13.

вует о естественном распадении совокупности исходных признаков на две группы: группу гуманитарных признаков ($x^{(1)}$, $x^{(2)}$, $x^{(3)}$) и группу математических признаков ($x^{(4)}$, $x^{(5)}$, $x^{(6)}$).

Подобная геометрическая интерпретация помогает выбрать вращение системы общих факторов, наиболее подходящее в отношении возможности их содержательной интерпретации. Дело в том, что, как уже отмечали, параметры модели факторного анализа, в том числе и сами общие факторы $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$, определяются не однозначно, а лишь с точностью до некоторого ортогонального преобразования, т. е. с точностью до вращения осей $f^{(1)}$, $f^{(2)}$, ..., $f^{(p')}$ в пространстве. При этом выбор окончательного решения, т. е. закрепление системы $f^{(1)}$, $f^{(2)}$, ..., $f^{(p')}$ в определенном положении, находится в распоряжении исследователя. Другими словами, исследователь должен решить вопрос: как, располагая некоторым частным решением $f^{(1)}$, $f^{(2)}$, ..., $f^{(p')}$, полученным, например, с помощью центроидного метода, выбрать такое ортогональное преобразование, такой поворот осей $f^{(1)}$, $f^{(2)}$, ..., $f^{(p')}$, при котором получаемые при этом новые общие факторы $\tilde{f}^{(1)}, \tilde{f}^{(2)}, \ldots, \tilde{f}^{(p')}$ допускают наиболее естественную и убедительную содержательную интерпретацию. Рассматривая расположение исходных признаков в плоскости $\hat{f}^{(1)}$ 0 $\hat{f}^{(2)}$ или в пространстве, натянутом на первые три

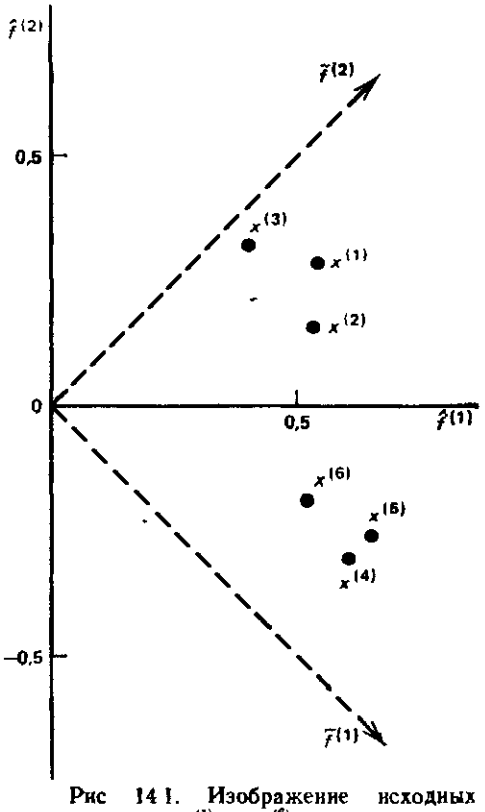


Рис 14.1. Изображение исходных признаков $x^{(1)}, \ldots, x^{(6)}$ в плоскости двух общих факторов $\hat{f}^{(1)}$, $\hat{f}^{(2)}$

общих фактора, естественно повернуть координатную систему таким образом, чтобы координатные оси прошли через наиболее четко выраженные сгущения точек-признаков (см. поворот, намеченный пунктирными осями $\tilde{f}^{(1)}$ и $\tilde{f}^{(2)}$ на рис. 14.1). При этом иногда бывает полезно отказаться от ортогональности общих факторов, переходя к косоугольной системе координат.

14.3. Некоторые эвристические методы снижения размерности

14.3.1. Природа эвристических методов. Описанные выше методы сокращения размерности исследуемого признакового пространства (метод главных компонент и модели факторного анализа) допускали интерпретацию в терминах той или иной строгой вероятностной модели и, следовательно, подразумевали возможность исследования свойств рассматриваемых процедур в рамках теории математической статистики. В данном параграфе речь пойдет о методах, подчиненных некоторым частным целевым установкам (наименьшее искажение геометрической структуры исходных «выборочных точек», наименьшее искажение их эталонного разбиения на классы и т. д.), но не формулируемых в терминах вероятностно-статистической теории[1]. Процедура выбора целевой установки, подходящей именно для данной конкретной задачи, практически не формализована, носит эвристический характер, т. е., как правило, обусловливается лишь опытом и интуицией исследователя. Поэтому будем называть такие методы *эвристическими*.

При отсутствии априорной или выборочной предварительной информации о природе исследуемого вектора наблюдений и о генеральных совокупностях, из которых эти наблюдения извлекаются, точно в таком же невыгодном положении находятся и методы факторного анализа и главных компонент. Однако для них все-таки существует принципиальная возможность теоретического обоснования (при наличии соответствующей дополнительной информации), в то время как лишь некоторые из эвристических методов удает-

[1] Отсутствие строгой вероятностно-статистической модели, лежащей в основе тех или иных методов, не исключает возможности использования отдельных вероятностно-статистических понятий и соответствующей терминологии, как это имеет место, например, в методе экстремальной группировки признаков, в методе корреляционных плеяд и некоторых других.

ся впоследствии теоретически обосновать в рамках строгой математической модели.

Подчеркнем, что факт описания здесь методов снижения размерности, не использующих предварительной информации, например обучающих или квазиобучающих выборок, целесообразно расценивать лишь как следствие признания неизбежности ситуаций, в которых такой информации не имеется, но не как стремление рекламировать эти методы в качестве наиболее эффективных. В действительности же обоснование и эффективное решение задач снижения размерности без слепой надежды на удачу можно, по нашему мнению, получить лишь на пути глубокого профессионального анализа, дополненного статистическими методами, использующими предварительную выборочную (обучающую) информацию.

14.3.2. Метод экстремальной группировки признаков. При изучении сложных объектов, заданных многими параметрами, возникает задача разбиения параметров на группы, каждая из которых характеризует объект с какой-либо одной стороны. Но получение легко интерпретируемых результатов осложняется тем, что во многих приложениях измеряемые параметры (признаки) лишь косвенно отражают существенные свойства, которыми характеризуется данный объект.

Так, в психологии измеряемые параметры — это реакции людей на различные тесты, а выражением существенных свойств, общими факторами, являются такие характеристики, как тип нервной системы, работоспособность и т.д. Подобная природа формирования набора частных характеристик объекта или системы присуща широкому классу явлений и процессов в экономике, социологии, медицине, педагогике и т. д.

Оказывается, что во многих случаях изменение какого-либо общего фактора сказывается неодинаково на измеряемых признаках, в частности, исходная совокупность из p признаков обнаруживает такое естественное «расщепление» на сравнительно (с p) небольшое количество групп, при котором изменение признаков, относящихся к какой-либо одной группе, обусловливается в основном каким-то одним общим фактором, своим для каждой такой группы. После принятия этой гипотезы разбиение на группы естественно строить так, чтобы параметры, принадлежащие к одной группе, были коррелированы сравнительно сильно, а параметры, принадлежащие к разным группам, — слабо. После такого разбиения для каждой группы признаков строится случайная величина, которая в некотором смысле наиболее сильно коррелирована с параметрами данной группы; эта случай-

ная величина интерпретируется как искомый фактор, от которого существенно зависят все параметры данной группы.

Очевидно, подобная схема является одним из частных случаев общей логической схемы факторного анализа. В отличие от ранее описанных классических моделей факторного анализа при эвристически-оптимизационном подходе группировка признаков и выделение общих факторов делаются на основе экстремизации некоторых эвристически введенных функционалов. Разбиения, оптимизирующие функционал J_1 или J_2 (см. ниже), называются *экстремальной группировкой параметров*. Вообще под задачей экстремальной группировки набора случайных величин $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ на заранее заданное число классов p' понимают отыскание такого набора подмножеств $S_1, S_2, \ldots, S_{p'}$ натурального ряда чисел $1, 2, \ldots, p$, что $\bigcup_{l=1}^{p'} S_l = \{1, 2, \ldots, p\}$, а $S_l \cap S_q = \varnothing$ при $l \neq q$, и таких p' нормированных (т. е. с единичной дисперсией $\mathrm{D}f^{(l)} = 1$) факторов $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$, которые максимизируют какой-либо критерии оптимальности.

Остановимся здесь на алгоритмах для двух различных критериев оптимальности [33].

Первый алгоритм экстремальной группировки признаков в качестве критерия оптимальности использует функционал

$$J_1 = \sum_{i \in S_1} [\mathrm{cor}\,(x^{(i)}, f^{(1)}]^2 + \ldots + \sum_{i \in S_{p'}} [\mathrm{cor}\,(x^{(i)}, f^{(p')})]^2,$$

в котором под $\mathrm{cor}\,(x, f)$ понимается обычный парный коэффициент корреляции между признаком x и фактором f. Обозначим $A_l = \{x^{(i)}, i \in S_l\}$, $l = 1, 2, \ldots, p'$. Максимизация функционала J_1 (как по разбиению признаков на группы $A_1, \ldots A_{p'}$, так и по выбору факторов $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$) отвечает требованию такого разбиения параметров, когда в одной группе оказываются наиболее «близкие» между собой, в смысле степени коррелированности, признаки: в самом деле, при максимизации функционала J_1, для каждого фиксированного набора случайных величин $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$, в одну l-ю группу будут попадать такие признаки, которые наиболее сильно коррелированы с величиной $f^{(l)}$; в то же время среди всех возможных наборов случайных величин $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$ будет выбираться такой набор, что каждая из величин $f^{(l)}$ в среднем наиболее «близка» ко всем признакам своей группы.

Очевидно, что при заданных классах $S_1, S_2, \ldots, S_{p'}$ оптимальный набор факторов $f^{(1)}, f^{(2)}, \ldots, f^{(p')}$ получается

в результате независимой максимизации каждого слагаемого

$$\sum_{i \in S_l} [\mathrm{cor}\,(x^{(i)},\ f^{(l)})]^2 \quad (l = \overline{1,\ p'}),$$

откуда

$$\max_{f^{(1)},\ f^{(2)},\ \ldots,\ f^{(l)}} J_1 = \sum_{l=1}^{p'} \lambda_l^2,$$

где λ_l — максимальное собственное значение матрицы Σ_l, составленной из коэффициентов корреляции переменных, входящих в A_l. При этом оптимальный набор факторов $f^{(l)}$, $l = 1,2, \ldots, p'$ задается формулами:

$$f^{(l)} = \frac{\sum\limits_{i \in S_l} \alpha_i^{(l)} x^{(i)}}{\sqrt{\sum\limits_{i,\ j \in S_l} \alpha_i^{(l)} \alpha_j^{(l)} r_{ij}}}, \quad l = 1,\ 2,\ \ldots,\ p', \qquad (14.14)$$

где $r_{ij} = \mathrm{cor}\,(x^{(i)}, x^{(j)})$, а $\alpha^{(l)} = (\alpha_1^{(l)}, \alpha_2^{(l)}, \ldots, \alpha_{m_l}^{(l)})$ — собственный вектор матрицы Σ_l, отвечающий максимальному собственному значению λ_l, т. е. $\Sigma_l \cdot \alpha^{(l)} = \lambda_l \cdot \alpha^{(l)}$.

С другой стороны, считая известными факторы $f^{(1)}$, $f^{(2)}$, ..., $f^{(p')}$, нетрудно построить разбиение S_1, S_2, ..., $S_{p'}$, максимизирующее J_1 при фиксированных $f^{(1)}$, $f^{(2)}$, ... $f^{(p')}$, а именно:

$$S_l = \{i : \mathrm{cor}^2\,(x^{(i)},\ f^{(l)}) \geqslant \mathrm{cor}^2\,(x^{(i)},\ f^{(q)}) \text{ для всех } q = 1,\ 2,\ \ldots,\ p'\}. \qquad (14.15)$$

Соотношения (14.14) и (14.15) являются необходимыми условиями максимума J_1.

Для одновременного нахождения оптимального разбиения S_1, S_2, ..., $S_{p'}$ и оптимального набора факторов $f^{(1)}$, $f^{(2)}$, ..., $f^{(p')}$ предлагается итерационный алгоритм, последовательно осуществляющий выбор оптимальных (по отношению к разбиению, полученному на предыдущем шаге), факторов, а затем выбор разбиения, оптимального к факторам, полученным на предыдущем шаге.

Пусть на ν-м шаге итерации построено разбиение параметров на группы A_1, ..., $A_{p'}$. Для каждой такой группы параметров строят факторы $f_\nu^{(l)}$ по формуле (14.14) и новое $(\nu + 1)$ разбиение параметров $A_1^{(\nu+1)}$, ..., $A_{p'}^{(\nu+1)}$ в соответ-

ствии с правилом: параметр $x^{(i)}$ относится к группе $A_l^{(\nu+1)}$, если

$$\operatorname{cor}^2\left(x^{(i)},\ f_\nu^{(l)}\right) \geqslant \operatorname{cor}^2\left(x^{(i)},\ f_\nu^{(q)}\right)\ (q=1,\ 2,\ \ldots,\ p'). \qquad (14.16)$$

Если для некоторого параметра $x^{(i)}$ найдутся два или более факторов таких, что для $x^{(i)}$ и этих факторов в (14.16) имеет место равенство, то параметр $x^{(i)}$ относится к одной из соответствующих групп произвольно.

Очевидно, что на каждом шаге итераций функционал J_1 не убывает, поэтому данный алгоритм будет сходиться к максимуму. Максимум может быть локальным.

Для описания второго алгоритма экстремальной группировки признаков введем функционал

$$J_2 = \sum_{i \in S_1} |\operatorname{cor}(x^{(i)},\ f^{(1)})| + \sum_{i \in S_2} |\operatorname{cor}(x^{(i)},\ f^{(2)})| + \ldots +$$

$$+ \sum_{i \in S_{p'}} |\operatorname{cor}(x^{(i)},\ f^{(p')})|.$$

В содержательном смысле функционал J_2 похож на функционал J_1 и его максимизация также соответствует основному требованию к характеру разбиения признаков на группы. В [33] показано, что имеет место следующее утверждение. Необходимыми и достаточными условиями максимума функционала J_2 являются следующие:

разбиение параметров на группы A_1, ..., $A_{p'}$ таково, что функционал

$$J_3 = \sum_{l=1}^{p'} \sqrt{\mathbf{D}\left(\sum_{i \in S_l} g_i x^{(i)}\right)}$$

(где g_i — некоторые числовые коэффициенты, равные либо $+1$, либо -1) достигает максимума как по разбиению на группы, так и по значениям коэффициентов g_i. Здесь под $\mathbf{D}z$ понимается, как обычно, дисперсия случайной величины z;

факторы $f^{(l)}$ определяются соотношениями

$$f^{(l)} = \frac{\sum_{i \in S_l} g_i x^{(i)}}{\sqrt{\sum_{i,j \in S_l} g_i g_j r_{ij}}}. \qquad (14.17)$$

Логическая схема доказательства этого следующая. Сначала, варьируя функционал J_2 и используя метод множите-

лей Лагранжа для учета условия $\mathbf{D}f^{(l)} = 1$, показывают, что в точке максимума функционала J_2 фактор $f^{(l)}$ имеет вид (14.17). Затем доказывается, что если $f^{(l)}$ имеет вид (14.17), то при любом наборе коэффициентов $g_i = \pm 1$ и любом разбиении параметров на группы имеет место соотношение $J_2 \geqslant J_3$, а если же J_3 достигает максимума, то $J_2 = J_3$. Из этого утверждения следует, в частности, что для нахождения групп S_l и факторов $f^{(l)}$ достаточно максимизировать функционал J_3. При фиксированном разбиении на группы функционал J_3 достигает максимума тогда, когда для каждого l соответствующие коэффициенты g_i максимизируют величину

$$\mathbf{D}\left(\sum_{i \in S_l} g_i x^{(i)}\right). \tag{14.18}$$

Поэтому естественно воспользоваться рекуррентной процедурой максимизации J_3. В процедуре циклически перебираются переменные $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, и на каждом шаге принимается решение об отнесении очередного параметра $x^{(i)}$ к одной из групп $A_1, \ldots, A_{p'}$ и определяется знак g_i.

Пусть к ν-му шагу алгоритма построены разбиения параметров на группы $A_1^{(\nu)}, \ldots, A_{p'}^{(\nu)}$, вычислены коэффициенты $g_1^{(\nu)}, g_2^{(\nu)}, \ldots, g_{p'}^{(\nu)}$, равные $+$ или -1, и пусть на этом шаге рассматривается признак $x^{(i)} \in A_l^{(\nu)}$. Тогда строятся p' вспомогательных коэффициентов $g_{i,l}^{(\nu+1)}$ $(l = 1, \ldots, p')$ по формуле

$$g_{i,l}^{(\nu+1)} = \operatorname{sign} \sum_{x^{(j)} \in A_l^{(\nu)}} g_j^{(\nu)} r_{ij} \quad (j \neq i),$$

где

$$\operatorname{sign} x = \begin{cases} 1 \text{ при } x > 0, \\ 0 \text{ при } x = 0, \\ -1 \text{ при } x < 0 \end{cases}$$

и для всех $l = 1, 2, \ldots, p'$ вычисляются разности

$$\Delta_l^{(\nu+1)} = \sqrt{\mathbf{D}\left(\sum_{\substack{x_j \in S_l^{(\nu)} \\ j \neq i}} g_j^{(\nu)} x^{(j)} + g_{i,l}^{(\nu+1)} x^{(i)}\right)} - \sqrt{\mathbf{D}\left(\sum_{\substack{x_j \in S_l^{(\nu)} \\ j \neq i}} g_j^{(\nu)} x^{(j)}\right)}.$$

Затем выбирается такой номер $l = l^*$, что

$$\Delta_{l^*}^{(\nu+1)} = \max_{1 \leqslant l \leqslant p'} \Delta_l^{(\nu+1)},$$

и признак $x^{(i)}$ исключается из группы A_l и присоединяется к группе A_{l^*}; остальные группы признаков на этом шаге не меняются. В результате получаем новое разбиение признаков — $A_1^{(\nu+1)}$, $A_2^{(\nu+1)}$, ..., $A_{p'}^{(\nu+1)}$. Новые значения коэффициентов $g_{i, l^*}^{(\nu+1)}$ определяются по формулам: $g_j^{(\nu+1)} = g_j^{(\nu)}$ (для $j \neq i$), $g_i^{(\nu+1)} = g_{i, l^*}^{(\nu+1)}$.

На следующем $(\nu + 1)$-м шаге алгоритма рассматривается параметр $x^{(i+1)}$, если $i \neq p$, и $x^{(1)}$, если $i = p$.

Процедура заканчивается, если при рассмотрении всех признаков очередного цикла сохранились как разбиения признаков на группы, так и значения всех коэффициентов; полученное разбиение и значения коэффициентов рассматриваются как оптимальные.

Для демонстрации сходимости метода к локальному максимуму в [33] доказывается, что на каждом шаге алгоритма значение J_3 не убывает.

Нетрудно проследить идейную близость метода экстремальной группировки факторов с методами, опирающимися на логическую схему факторного анализа. Так, например, отправляясь от общей модели вида

$$x^{(i)} = \sum_{q=1}^{p'} l_{iq} f^{(q)} + u_i$$

(14.1), первую компоненту $f^{(1)}$ и «нагрузки» l_{i1} в методе главных компонент можно определять из условия минимума выражения $\sum_{i=1}^{p} \mathbf{D}\,(x^{(i)} - l_{i1} f^{(1)})$ при нормирующем ограничении $\mathbf{D} f^{(1)} = 1$. Решение этой условно экстремальной задачи очевидным образом сводится к нахождению максимума выражения $\sum_{i=1}^{p} [\mathrm{cor}\ (x^{(i)}, f^{(1)})]^2$ при условии $\mathbf{D} f^{(1)} = 1$.

Для построения следующего фактора $f^{(2)}$ (второй главной компоненты) рассматриваются случайные величины $x^{(i2)} = x^{(i)} - \mathrm{cor}\ (x^{(i)}, f^{(1)})\, f^{(1)}$. Для этих случайных величин аналогичным образом находится свой фактор, который и является фактором $f^{(2)}$, и т. д.

Очевидно, что при реализации первого алгоритма метода экстремальной группировки признаков для каждой группы признаков A_l строится фактор, имеющий смысл первой главной компоненты для признаков этой группы.

В центроидном методе общий фактор ищут в виде

$$f^{(1)} = \sum_{i=1}^{p} g_i x^{(i)}, \tag{14.19}$$

где $g_i = \pm 1$ и g_i выбирается так, чтобы максимизировать величину

$$\mathbf{D}f^{(1)} = \mathbf{D}\left(\sum_{i=1}^{p} g_i x^{(i)} \right). \tag{14.20}$$

Сравнение выражений (14.19) и (14.20) с выражениями (14.17) и (14.18) показывает, что максимизация функционала J_2 приводит к построению для каждой группы признаков фактора, отличающегося на некоторый множитель от первого общего фактора, который был бы построен для этой группы центроидным методом.

14.3.3. Метод корреляционных плеяд. Задача разбиения признаков на группы часто имеет и самостоятельное значение. Например, в ботанике для систематизации вновь открытых растений делают разбиение набора признаков на группы так, чтобы 1-я группа характеризовала форму листа, 2-я группа — форму плода и т. д. В связи с этим и возник эвристический метод корреляционных плеяд [48, 151].

Метод корреляционных плеяд, так же как и метод экстремальной группировки, предназначен для нахождения таких групп признаков—«плеяд», когда корреляционная связь, т. е. сумма модулей коэффициентов корреляции между параметрами одной группы (внутриплеядная связь) достаточно велика, а связь между параметрами из разных групп (межплеядная) — мала. По определенному правилу по корреляционной матрице признаков образуют чертеж — граф, который затем с помощью различных приемов разбивают на подграфы. Элементы, соответствующие каждому из подграфов, и образуют плеяду.

Рассмотрим корреляционную матрицу $\mathbf{R} = (r_{ij})$, $i, j = 1, 2, \ldots, p$, исходных признаков. Нарисуем p кружков; внутри каждого кружка напишем номер одного из признаков. Каждый кружок соединяется линиями со всеми остальными кружками; над линией, соединяющей i-й и j-й элементы (ребром графа), ставится значение модуля коэффициента корреляции $|r_{ij}|$. Полученный таким образом чертеж рассматриваем как исходный граф.

Задавшись (произвольным образом или на основании предварительного изучения корреляционной матрицы) некоторым пороговым значением коэффициента корреляции

r_0, исключаем из графа все ребра, которые соответствуют коэффициентам корреляции, по модулю меньшим r_0. Затем задаем некоторое $r_1 > r_0$ и относительно него повторяем описанную процедуру. При некотором достаточно большом r граф распадается на несколько подграфов, т. е. таких групп кружков, что связи (ребра графа) между кружками различных групп отсутствуют. Очевидно, что для полученных таким образом плеяд внутриплеядные коэффициенты корреляции будут больше r, а межплеядные — меньше r.

В другом варианте корреляционных плеяд [48] предлагается упорядочивать признаки и рассматривать только те коэффициенты корреляции, которые соответствуют связям между элементами в упорядоченной системе.

Упорядочение производится на основании принципа максимального корреляционного пути: все p признаков связываются при помощи $(p-1)$ линий (ребер) так, чтобы сумма модулей коэффициентов корреляции была максимальной. Это достигается следующим образом: в корреляционной матрице находят наибольший по абсолютной величине коэффициент корреляции, например $|r_{l,m}| = r^{(1)}$ (коэффициенты на главной диагонали матрицы, равные единице, не рассматриваются).

Рисуем кружки, соответствующие параметрам $x^{(l)}$ и $x^{(m)}$ и над «связью» между ними пишем значение $r^{(1)}$. Затем, исключив $r^{(1)}$, находим наибольший коэффициент в m-м столбце матрицы (это соответствует нахождению признака, который наиболее сильно после $x^{(l)}$ «связан» с $x^{(m)}$, и наибольший коэффициент в l-й строке матрицы (это соответствует нахождению признака, наиболее сильно после $x^{(m)}$ «связанного» с $x^{(l)}$). Из найденных таким образом двух коэффициентов корреляции выбирается наибольший — пусть это будет $|r_{lj}| = r^{(2)}$. Рисуем кружок $x^{(j)}$, соединяем его с кружком $x^{(l)}$ и проставляем значение $r^{(2)}$. Затем находим признаки, наиболее связанные с $x^{(l)}$, $x^{(m)}$ и $x^{(j)}$, и выбираем из найденных коэффициентов корреляции наибольший. Пусть это будет $|r_{jq}| = r^{(3)}$. Требуем, чтобы на каждом шаге получался новый признак, поэтому признаки, уже изображенные на чертеже, исключаются, следовательно, $q \neq l$, $q \neq m$, $q \neq j$.

Далее рисуем кружок, соответствующий $x^{(q)}$, и соединяем его с $x^{(j)}$ и т. д. На каждом шаге находятся параметры, наиболее сильно связанные с двумя последними рассмотренными параметрами, а затем выбирается один из них, соответствующий бо́льшему коэффициенту корреляции. Процедура заканчивается после $(p-1)$-го шага; граф оказывается состоящим из p кружков, соединенных $(p-1)$ ребром. Затем задается пороговое значение r, а все ребра, соответству-

ющие меньшим, чем r, коэффициентам корреляции, исключаются из графа.

Назовем *незамкнутым графом* такой граф, для которого для любых двух кружков существует единственная траектория, составленная из линий связи, соединяющая эти два кружка. Очевидно, что во втором варианте метода корреляционных плеяд допускается построение только незамкнутых графов, а в первом варианте такое ограничение отсутствует. Поэтому разбиения на плеяды, полученные разными способами, могут не совпадать.

В работе [97] приводятся результаты экспериментальной проверки алгоритмов экстремальной группировки параметров, а также сравнение полученных результатов с результатами, даваемыми методом корреляционных плеяд.

Эксперимент проводился на физиологическом материале: исследовались влияния шумов и вибрации на работоспособность и самочувствие. Регистрировались 33 признака ($p = 33$), из них 7 параметров, характеризующих температуру тела; 4 — кровяное давление; 14 — аудиометрию (порог слышимости на заданной частоте); 2 — дыхание; 4 — силу и выносливость рук и 2 (особенных параметра) — пульс и скорость реакции.

С точки зрения физиолога «идеальным» было бы разбиение, при котором все характеристики температур образовали бы отдельную группу; параметры, характеризующие давление, — свою отдельную группу и т.д., обособленные параметры образовали бы группы, состоящие из одного элемента. Наиболее близким к «идеальному» оказалось разбиение, полученное вторым алгоритмом экстремальной группировки, хотя алгоритм и присоединяет обособленные параметры к другим группам. Наименее точные (среди трех сравниваемых алгоритмов) результаты дал метод корреляционных плеяд.

Исторически раньше возникшие различные варианты метода корреляционных плеяд являются в действительности несколько упрощенными эвристическими версиями более совершенных в математическом плане алгоритмов исследования структуры связей между компонентами многомерного признака, использующими графы-деревья (см. [12, гл. 4]).

14.3.4. Снижение размерности с помощью кластер-процедур. В ряде ситуаций удобно рассматривать признаки $x^{(i)}$ ($i = 1,2, \ldots, p$) как одномерные наблюдения и использовать многократное повторение этих наблюдений (на n исследуемых объектах) для введения и вычисления таких естественных мер близости между объектами (признаками) $x^{(i)}$ и $x^{(j)}$,

какими являются в данном случае абсолютная величина коэффициента корреляции r_{ij} или корреляционное отношение η_{ij} (вычисления r_{ij} и η_{ij} и их свойства см., например, [12])[1].

Следуя идее обобщенного (степенного) среднего (см. гл. 5), введем в качестве меры близости групп признаков A_l и A_q величину

$$R_{lq}^{(\tau)} = \left[\frac{1}{m_l m_q} \sum_{x^{(i)} \in A_l} \sum_{x^{(j)} \in A_q} |r_{ij}|^{\tau} \right]^{\frac{1}{\tau}},$$

где τ — некоторый числовой параметр, выбор конкретного значения которого находится в нашем распоряжении; m_v — число признаков, составляющих группу A_v. Аналогично вводится средняя мера близости $R(A_l)$ признаков, входящих в одну группу

$$R(A_l) = R_{ll}^{(\tau)} = \left(\frac{1}{m_l^2} \sum_{x^{(i)} \in A_l} \sum_{x^{(j)} \in A_l} |r_{ij}|^{\tau} \right)^{\frac{1}{\tau}}.$$

Если желаемая размерность p' ($p' < p$) задана заранее, то исходные p признаков $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ разбивают на p' однородных групп одним из двух способов: либо, последовательно объединяя в одну группу два наиболее близких, в смысле r_{ij} или $R_{lq}^{(\tau)}$, признака (или признак и группу, или две группы) до тех пор, пока не останется ровно p' групп (иерархическая кластер-процедура), либо, находя такое разбиение исходных признаков на p' групп, при котором усредненная мера внутригрупповой близости признаков $\overline{R^{(\tau)}} = \left(\frac{1}{p'} \sum_{l=1}^{p'} m_l |R(A_l)|^{\tau} \right)^{\frac{1}{\tau}}$ была бы максимальной. Последнего обычно удается добиться с помощью простого перебора вариантов, так как общее число признаков p, как правило, не

---

[1] Если для описания меры близости между $x^{(i)}$ и $x^{(j)}$ используется корреляционное отношение, то предварительно целесообразно произвести симметризацию этой меры, рассматривая в качестве симметричной характеристики степени близости этих признаков величину $\tilde{\eta}_{ij} = \frac{1}{2}(\eta_{ij} + \eta_{ji})$, где η_{ij} — обычное корреляционное отношение переменной $x^{(i)}$ по переменной $x^{(j)}$.

превосходит несколько десятков, а p' — несколько единиц. После этого от каждой группы следует отобрать по одному представителю, используя для этого технику метода главных компонент или факторного анализа (отдельно внутри каждой группы).

Если желаемая размерность p' заранее не определена, то разбиение исходных признаков на группы, а следовательно, и выбор неизвестного p' можно подчинить задаче максимизации функционала типа $\bar{R}^{(\tau)} + Z_\tau$, где Z_τ — введенная в гл. 5 мера концентрации разбиения, т. е.

$$Z_\tau = \left[\frac{1}{p} \sum_{i=1}^{p} \left(\frac{v(x^{(i)})}{p} \right)^\tau \right]^{\frac{1}{\tau}}.$$

Здесь $v(x^{(i)})$ — число признаков в группе, содержащей признак $x^{(i)}$. Можно воспользоваться также и двойственной формулировкой экстремальной задачи разбиения объектов (признаков) на неизвестное число групп.

ВЫВОДЫ

1. Различные версии моделей и методов факторного анализа (центроидный, максимального правдоподобия, экстремальной группировки параметров, корреляционных плеяд и др.) основаны на общей базовой идее, в соответствии с которой значения всех признаков $x^{(1)}, \ldots, x^{(p)}$ анализируемого набора формируются под воздействием сравнительно небольшого числа одних и тех же (общих) факторов $f^{(1)}, \ldots, f^{(p')}$, не поддающихся, правда, непосредственному измерению (и потому называющихся латентными). В определенном смысле эти общие факторы выступают в роли причин, а наблюдаемые (анализируемые) признаки — в роли следствий.
2. Поскольку число общих (латентных) факторов существенно меньше числа анализируемых признаков, то методы факторного анализа в конечном счете нацелены (так же как и метод главных компонент) *на снижение размерности анализируемого признакового пространства*.
3. Статистическая реализация модели факторного анализа предусматривает последовательное решение вопросов *существования* такой модели, ее *идентификации* (т. е. возможности ее однозначного восстановления по исход-

ным статистическим данным), *алгоритмического определения ее структурных параметров* (т. е. определения способа вычисления неизвестных параметров модели при точно известной ковариационной матрице анализируемого многомерного признака) и *их статистической оценки по имеющимся наблюдениям*, включая статистические оценки для самих общих (латентных) факторов.

4. Наиболее распространенной в практике статистических исследований и наиболее теоретически разработанной является *каноническая модель* факторного анализа, в которой признаки линейно зависят от факторов, факторы взаимно некоррелированы между собой и со случайными остатками модели, а случайные остатки в свою очередь взаимно некоррелированы и нормально распределены.

5. Между методом главных компонент и линейной моделью факторного анализа имеется идейная общность: и тот и другой метод можно рассматривать как метод аппроксимации набора анализируемых переменных с помощью линейных функций от сравнительно небольшого числа одних и тех же вспомогательных переменных (главных компонент — в одном методе и общих факторов — в другом). Их небольшое различие — лишь в конкретизации критерия точности аппроксимации.

6. Наиболее «узкие места» в практической дееспособности модели факторного анализа связаны с решением задачи оценки числа p' общих факторов модели и с содержательной интерпретацией найденных общих факторов. Для успешного решения последней задачи широко пользуются неоднозначностью (с точностью до ортогонального преобразования) определения общих факторов и соответственно возможностью их разнообразных «вращений» в факторном пространстве.

7. Наряду с математико-статистическими методами снижения размерности, т. е. с методами, допускающими описание и интерпретацию в терминах строгой вероятностной модели, существуют и широко используются в статистической практике так называемые *эвристические методы*. Свое название они оправдывают тем, что порождаются обычно некоторыми частными целевыми установками, выраженными в виде *установленных на содержательно-субъективном уровне* оптимизируемых критериев качества решения задачи. К таким методам, в частности, относятся методы экстремальной группировки параметров, метод корреляционных плеяд, некоторые «кластерные» приемы и т. п.

Глава 15. ЭКСПЕРТНО-СТАТИСТИЧЕСКИЙ МЕТОД ПОСТРОЕНИЯ ЕДИНОГО СВОДНОГО ПОКАЗАТЕЛЯ ЭФФЕКТИВНОСТИ ФУНКЦИОНИРОВАНИЯ (КАЧЕСТВА) ОБЪЕКТА (СКАЛЯРНАЯ РЕДУКЦИЯ МНОГОКРИТЕРИАЛЬНОЙ СХЕМЫ)

И в профессиональной деятельности, и в своей повседневной жизни человек постоянно сталкивается с ситуациями, когда ему приходится сравнивать между собой и упорядочивать по некоторому *не поддающемуся непосредственному измерению* свойству ряд объектов. Речь может идти, в частности, о сравнении стран по прогрессивности их макроструктуры потребления, предприятий отрасли по эффективности их деятельности, сложных изделий (например, определенного программного средства) по обобщенной характеристике качества, специалистов по эффективности их участия в выполнении поставленной задачи, участников игровых видов спорта по уровню проявленного ими (в определенном состязании) мастерства и т. д. Формализации подобных ситуаций и вытекающим из нее рекомендациям по построению некоторого условного измерителя упомянутого свойства объекта и посвящена данная глава.

15.1. Латентный единый (сводный) показатель «качества». Понятия «выходного качества» целевой функции и «входных переменных» (частных критериев)

Пусть обобщенная сводная характеристика f анализируемого свойства объекта определяется набором частных критериев, задаваемых поддающимися учету и измерению переменными $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ (в дальнейшем будем называть их «входными»), однако сама эта характеристика является *латентной*, т. е. *не поддается непосредственному количественному измерению* (для нее не существует объективно обусловленной шкалы). Естественно предположить, что интуитивное экспертное (профессиональное) восприятие этой характеристики (обозначим его y) можно представить как несколько искаженное значение $f(x^{(1)}, \ldots, x^{(p)})$, причем это искажение δ носит случайный характер и обусловлено как разрешающей способностью такого «измерительного прибора», каковым в данной схеме является эксперт, так и су-

ществованием ряда слабо влияющих на y, но не входящих в состав $X = (x^{(1)}, \ldots, x^{(p)})$ «входных переменных». Тогда модель, связывающая между собой интуитивное представление о сводном показателе качества (y), сам сводный показатель как функцию от X ($f(X)$) и случайную погрешность $\delta(X)$, может быть определена в виде

$$y = f(X) + \delta(X). \qquad (15.1)$$

Практически, не ограничивая общности данной схемы, можно принять естественные допущения относительно первых двух моментов остаточной случайной компоненты $\delta(X)$:

$$\mathbf{E}\delta(X) \equiv 0, \quad \mathbf{D}\delta(X) = \sigma^2(X) < \infty. \qquad (15.1')$$

Тогда, очевидно, обобщенная (сводная) характеристика $f(X)$ может интерпретироваться как регрессия y по X, и если бы в качестве исходной статистической информации располагали бы наряду со значениями $X_i = (x_i^{(1)}, \ldots, x_i^{(p)})$ и результатами регистрации соответствующих значений зависимой переменной y_i (i — номер наблюдения), то данная схема непосредственно сводилась бы к обычной модели регрессии (см. [12, гл. 5]). Специфика модели (15.1), (15.1′) состоит в том, что *вместо прямых измерений y можно получить (с помощью экспертов) лишь некоторые специального вида сведения о его значениях*, чаще всего — сравнительного плана (типа ранжировок или парных сравнений обследованных объектов по свойству y). Это обусловливает и более скромные претензии в отношении целей статистического анализа модели (15.1), (15.1′): вместо требуемого в регрессионном анализе восстановления (оценивания) функции $f(X)$ ставится задача оценивания $f(X)$ с точностью до произвольного монотонного преобразования.

Определение. *Целевой функцией* исследуемого обобщенного свойства («выходного качества») называется любое преобразование вида $\varphi(x^{(1)}, \ldots x^{(p)}) = \varphi(X)$, сохраняющее заданное соотношение порядка между анализируемыми объектами $O_1, O_2, \ldots, O_n$ по усредненным значениям выходного качества, т. е. обладающее тем свойством, что из $f(X_{i_1}) \geqslant f(X_{i_2}) \geqslant \ldots \geqslant f(X_{i_n})$ с необходимостью следует выполнение неравенств $\varphi(X_{i_1}) \geqslant \varphi(X_{i_2}) \geqslant \ldots \geqslant \varphi(X_{i_n})$, и, наоборот, из последней серии неравенств вытекает выполнение соответствующих неравенств для $f(X_{i_k})$, $k = 1, 2, \ldots, \ldots, n$. Очевидно, данное здесь определение целевой функции неоднозначно. Действительно, если $\varphi(X)$ есть целевая функция и $U(\varphi)$ — любая взаимно-однозначная монотонно возрастающая функция, то всякая функция вида

$\psi(X) = U[\varphi(X)]$ также будет целевой функцией. Это означает, что допущение о наличии определенной шкалы в измерении единого сводного показателя играет во многих случаях чисто вспомогательную роль и нацеливает на поиск, связанный с выявлением этой шкалы лишь с точностью до произвольного допустимого преобразования шкал. Ведь в соответствии с данным определением само значение целевой функции не отражает никакой реальной, физически содержательной количественной закономерности. Реальные закономерности отражаются только соотношениями «больше» или «меньше» между значениями этой функции для различных наборов величин входных параметров $X = (x^{(1)}, \ldots, x^{(p)})$. Тем самым эти соотношения отражают предпочтение с точки зрения анализируемого выходного качества одних значений X перед другими. Поэтому в задачах, в которых возможно регулирование значений X (в некоторой допустимой области), наиболее рациональным управлением естественно признать то, которое максимизирует, при заданных ограничениях на X, значения целевой функции.

Данное определение целевой функции допускает ее *содержательную* (экономическую, социально-экономическую, квалиметрическую, психологическую и т.д.) интерпретацию. Помимо приведенных ниже примеров оно может быть использовано при построении и анализе различных глобальных и частных целевых функций благосостояния и потребления (о которых речь идет, например, в [52, 99, 188, 247], в других задачах аналогичного профиля.

Итак, функция $f(X)$, с помощью которой можно было бы производить сравнительную оценку анализируемого «выходного качества» на рассматриваемых объектах, определена лишь с точностью до произвольного монотонного преобразования. Тем не менее для построения алгоритма ее восстановления было бы удобно параметризовать модель (15.1), т. е. определить параметрическое семейство $\mathbf{F} = \{f(X; \Theta)\}$, в рамках которого будет производиться поиск целевой функции $f(X)$. Выбор этого параметрического семейства, как правило, не удается подкрепить исчерпывающим теоретическим обоснованием, а потому с этого момента исследователь имеет дело не с целевой функцией $f(X)$, а с некоторой ее аппроксимацией $\widehat{f}(X)$. Это не должно смущать. Оперирование с аппроксимацией избавляет от необходимости постулирования существования самой целевой функции (что в ряде ситуаций является весьма спорным моментом): в то время как сама целевая функция как объективно существующая универсальная скалярная характеристика выходного

качества может и не существовать, ее аппроксимация имеет определенный условный смысл и может плодотворно использоваться как некая вспомогательная характеристика в ограниченном интервале времени и при некоторых заранее оговоренных условиях.

Имея в виду достаточную однородность обследуемых объектов по всем неучтенным переменным, т. е. по переменным, не вошедшим в состав $x^{(1)}, \ldots, x^{(p)}$, и ограниченность интервала времени, в течение которого будем использовать искомую аппроксимацию целевой функции, а также реализуя идею разложения любой функции в ряд Тейлора, ограничимся в дальнейшем изложении аппроксимациями линейного и квадратичного вида, т. е.

$$\widehat{f}(X;\ \Theta) = \sum_{i=1}^{p} \theta_i x^{(i)}\ ^1$$

и

$$\widehat{f}(X;\ \Theta) = \sum_{i=1}^{p} \theta_i x^{(i)} + \sum_{i,\ j=1}^{p} \theta_{ij} x^{(i)} x^{(j)}.$$

Коэффициенты θ_i и θ_{ij} оцениваются статистически (см. § 15.3) по исходным данным, структура и происхождение которых описываются в следующем параграфе.

15.2. Исходные данные

Итак, пусть речь идет о построении непосредственно не поддающегося измерению единого сводного показателя эффективности функционирования (качества) объекта и пусть с этой целью были собраны исходные данные по n таким объектам: $O_1, O_2, \ldots, O_n$. На основании этих исходных данных как раз и оцениваются параметры Θ искомой целевой функции $f(X; \Theta)$. Эти исходные данные состоят из двух частей: *экспертной* и *статистической* (отсюда название метода).

Экспертная часть исходных данных. Эта часть исходных данных относится к сведениям о значениях случайной величины y_i (i — номер обследованного объекта) в модели (15.1) и получается с помощью специально организованного опроса экспертов и соответствующей статистической обработки экспертных оценок. При этом сведения об y_i ($i =$

[1] При конкретизации постановки задачи коэффициентам θ_i и θ_{ij} часто удается дать содержательную интерпретацию [188].

$= 1,2, \ldots, n$) получают от экспертов в одной из следующих форм.

Форма (а) — наиболее информативный (а потому наиболее трудный для экспертов) вариант. Предусматривает получение *экспертных балльных оценок* выходного качества

$$\begin{aligned} &y_{11э}, y_{21э}, \ldots, y_{n1э} \\ &y_{12э}, y_{22э}, \ldots, y_{n2э} \\ &\ldots\ldots\ldots\ldots \\ &y_{1mэ}, y_{2mэ}, \ldots, y_{nmэ}, \end{aligned} \qquad (15.2а)$$

где $y_{ij_э}$ — оценка выходного качества объекта O_i, полученная от j-го эксперта (здесь n— число оцениваемых объектов, m — число участвующих в оценке экспертов).

Форма (б) — средний по информативности (и по степени трудности для экспертов) вариант.

Предусматривает получение лишь *экспертных упорядочений* обследованных объектов по степени проявления в них анализируемого свойства, т. е. ранжировок вида

$$\begin{aligned} &R_{11э}, R_{21э}, \ldots, R_{n1э} \\ &R_{12э}, R_{22э}, \ldots, R_{n2э} \\ &\ldots\ldots\ldots\ldots \\ &R_{1mэ}, R_{2mэ}, \ldots, R_{nmэ}, \end{aligned} \qquad (15.2б)$$

где $R_{ij_э}$ — ранг (место), присвоенный объекту O_i j-м экспертом в ряду из n обследованных объектов, упорядоченном этим экспертом по степени проявления анализируемого свойства.

Форма (в) — наименее информативный (и наименее трудный для экспертов) вариант. Информация от каждого (j-го) эксперта поступает в форме *булевой матрицы парных сравнений*

$$\begin{aligned} &\gamma_{j_э} = (\gamma_{ik.j_э}),\ i,\ k = \overline{1,\ n}, \\ &j = 1,\ 2,\ \ldots\ m. \end{aligned} \qquad (15.2в)$$

где $\gamma_{ik.\,j_э}$ — результат парного сравнения j-м экспертом объектов O_i и O_k, может выражаться либо единицей, либо нулем по одному из следующих правил:

если эксперт производит сравнение объектов O_i и O_k типа их упорядочения по анализируемому свойству, то

$$\gamma_{ik.j_э} = \begin{cases} 1, \text{ если по мнению } j\text{-го эксперта } O_i \\ \quad \text{не хуже } O_k, \\ 0 — \text{в противном случае;} \end{cases} \qquad (15.2в')$$

если эксперт производит сравнение объектов O_i и O_k лишь с точки зрения принадлежности этих объектов к однородному (по анализируемому свойству) классу, то

$$\gamma_{ik,j_0} = \begin{cases} 1, \text{ если объекты } O_i \text{ и } O_k \text{ однородны,} \\ 0 — \text{в противном случае.} \end{cases} \qquad (15.2в')$$

Вычислительные трудности, связанные с реализацией алгоритма оценивания параметров Θ искомой целевой функции $f(X; \Theta)$, естественно, возрастают по мере перехода от более информативных вариантов экспертной информации об y_i к менее информативным.

Статистическая часть исходных данных. Как выше уже отмечено, входные переменные (частные критерии) $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, на основании которых формируется представление об исследуемом выходном качестве, поддаются непосредственному измерению (регистрации) на каждом из обследуемых объектов. Поэтому статистически обследовав анализируемые объекты O_1, O_2, ..., O_n по переменным $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, будем иметь *статистическую* часть исходных данных в виде матрицы (таблицы) типа «объект — свойство»:

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(p)} & x_2^{(p)} & \dots & x_n^{(p)} \end{pmatrix}, \qquad (15.3)$$

где $x_i^{(l)}$ — значение l-й входной переменной, зарегистрированное на i-м объекте.

Таким образом, приступая к оценке параметров Θ искомой целевой функции $f(X; \Theta)$ в модели (15.1), исследователь располагает исходной информацией об объектах O_1, O_2, ..., O_n, состоящей из данных таблицы (15.3) и одного из вариантов (15.2а) — (15.2в) сведений об y_i.

15.3. Алгоритмические и вычислительные вопросы построения неизвестной целевой фуикции

15.3.1. Общая логическая схема оценивания параметров Θ целевой функции $f(X; \Theta)$. Располагая конкретными значениями Θ_0 параметров Θ, для каждого объекта O_i можем вычислить величину единого сводного показателя $f(X_i; \Theta_0)$ и далее, ориентируясь на сравнение значений $f(X_1; \Theta_0)$, $f(X_2; \Theta_0)$, ..., $f(X_n; \Theta_0)$, получить основанную на целевой

функции ранжировку объектов по искомому выходному качеству

$$R_1(\Theta_0),\ R_2(\Theta_0),\ \ldots,\ R_n(\Theta_0) \tag{15.4}$$

либо их разбиение на однородные (по f) классы, которое так же, как и уже имеющиеся экспертные разбиения (15.2в), может быть представлено в виде булевой матрицы $\gamma(\Theta_0; \Delta)$. (Способ получения такого разбиения с помощью вычисления значений целевой функции $f(X_i; \Theta_0)$ и смысл «свободного» скалярного параметра Δ объяснены ниже).

Оценку $\widehat{\Theta}$ неизвестных параметров Θ предлагается подбирать таким образом, чтобы: 1) минимизировать расхождения в экспертных ($y_{ij_\ni}$) и полученных с помощью целевой функции ($f(X_i; \Theta)$) балльных оценках выходного качества (в варианте (*а*) экспертной информации); 2) максимизировать согласованность экспертных и полученной с помощью целевой функции ранжировок объектов по анализируемому выходному качеству (в варианте (*б*) экспертной информации); 3) минимизировать расхождения в экспертных и полученном с помощью целевой функции разбиениях объектов на классы (в варианте (*в*) экспертной информации).

Из сказанного следует, что экспертно-статистический метод построения единого сводного показателя нацелен на формализацию (в виде соответствующим образом подобранной целевой функции $f(X; \Theta)$) тех критерийных установок, которыми руководствовались привлеченные к контрольному эксперименту эксперты при формировании своих оценок вида (15.2а), (15.2б) или (15.2в). *Поэтому состоятельность и эффективность этого метода целиком зависит от компетентности и согласованности используемых в нем экспертных оценок.*

15.3.2. Оценивание неизвестных параметров целевой функции при балльных экспертных оценках выходного качества.

В этом случае задача сводится к обычной схеме регрессионного анализа и соответственно к использованию метода наименьших квадратов. Действительно, располагая данными вида (15.2а) — (15.3), можем записать модель (15.1) в виде

$$\begin{cases} y_{ij_\ni} = f(X_i;\ \Theta) + \delta_j(X_i); \\ \mathbf{E}\delta_j(X_i) \equiv 0,\ \mathbf{D}\delta_j(X_i) = \sigma^2_{ij}, \end{cases} \tag{15.5}$$

где величина (как правило, неизвестная) σ^2_{ij} характеризует погрешность в оценке j-м экспертом выходного качества i-го

объекта. Критерий метода наименьших квадратов дает нам оценку $\widehat{\Theta}$ векторного параметра Θ как решение оптимизационной задачи вида

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{1}{\sigma^2_{ij}}(y_{ij_\ni}-f(X_i;\ \Theta))^2\to\min_{\Theta}. \qquad (15.6)$$

Если не располагают никакими сведениями относительно величин σ^2_{ij}, то принимают упрощающее предположение $\sigma^2_{ij}=\text{const}$, и задача (15.6) соответственно упрощается. В некоторых случаях удается априори задаться «весами» ω_j, оценивающими сравнительную компетентность j-го эксперта ($j=1,2,\ \ldots,\ m$). Тогда эти веса вставляются в (15.5) в качестве сомножителей слагаемых вместо величин $1/\sigma^2_{ij}$.

Конкретные рекомендации по вычислительной реализации решения задач типа (15.5) приведены в [12, гл. 7—9].

15.3.3. Оценивание неизвестных параметров целевой функции при экспертных ранжировках и парных сравнениях объектов. Каждая экспертная ранжировка $R_{\cdot j_\ni}=(R_{1j_\ni},\ R_{2j_\ni},\ \ldots,\ R_{nj_\ni})$, ($j$-я строка в (15.2б)) может быть представлена в виде булевой матрицы $\gamma_{j_\ni}$ в соответствии с правилом (15.2в′). Поэтому в дальнейшем (в данном пункте) будем считать, что экспертная информация о выходном качестве объектов представлена в виде матриц парных сравнений вида (15.2 в).

В общем случае задача состоит в том, чтобы на основе известных сравнений N пар объектов (*не обязательно всех возможных пар из n объектов*, т. е. N может быть меньше C_n^2) определить скаляриую функцию $f(X;\ \Theta)$, такую, что парные сравнения, установленные по этой функции отиосительно тех же пар объектов, минимально (в смысле заданного критерия) отличались бы от экспертно установленных.

В случае парных сравнений в виде отношений предпочтения (см. правило (15.2в′) формирования элементов $\gamma_{i_k\cdot j_\ni}$), поставив на первое место в каждой из N экспертно оцененных пар лучший (не худший) объект, будем иметь пары $(i_1,\ k_1)$, $(i_2,\ k_2)$, ..., $(i_N,\ k_N)$, значения целевых функций элементов которых должны были бы удовлетворять системе неравенств

$$\begin{cases} f(X_{i_q};\ \Theta)-f(X_{k_q};\ \Theta)\geqslant 0; \\ q=1,\ 2,\ \ldots,\ N. \end{cases} \qquad (15.7)$$

Однако в общем случае эта система оказывается несовместной. Поэтому в каждое неравенство (i_q, k_q) вводится невязка

$$b_{i_q k_q}(\Theta) = \begin{cases} 0, \text{ если } f(X_{i_q}; \Theta) - f(X_{k_q}; \Theta) \geqslant 0; \\ -(f(X_{i_q}; \Theta) - f(X_{k_q}; \Theta)) \text{ — в противном случае} \end{cases}$$

и вектор оценок $\widehat{\Theta}$ определяется из условия минимума суммы невязок $\sum_{q=1}^{N} b_{i_q k_q}(\Theta) \to \min_{\Theta}$ при некоторых ограничениях (типа нормировки) на компоненты искомого параметра Θ.

В работе [87] подробно расписан алгоритм вычисления оценки $\widehat{\Theta}$ для случая линейной целевой функции [1].

В случае парных сравнений, задающих разбиение объектов на однородные классы (см. правило (15.2в'') формирования элементов $\gamma_{ik,js}$), матрицы $\gamma_{1s}, \ldots, \gamma_{ms}$ задают m различных разбиений множества $\{O_1, O_2, \ldots, O_n\}$ на классы, элементы каждого из которых близки по анализируемому выходному качеству. Для любых двух разбиений γ_s и γ_r может быть введена мера близости между этими разбиениями

$$d(\gamma_s, \gamma_r) = \frac{1}{2} \sum_{i,j=1}^{n} |\gamma_{ij,s} - \gamma_{ij,r}|.$$

Пусть $\widehat{f}(X; \Theta) = \sum_{l=1}^{p} \theta_l x^{(l)}$ — некоторая линейная аппроксимация f. Задавшись некоторым $\varepsilon > 0$, можно с помощью $\widehat{f}(X; \Theta)$ построить разбиение n объектов на классы. В один класс при этом попадут те объекты, у которых $0 \leqslant \widehat{f}(X; \Theta) < \varepsilon$, в другой — те, у которых $\varepsilon \leqslant \widehat{f}(X; \Theta) < 2\varepsilon$ и т.д. Полученное разбиение $\gamma(\varepsilon, \Theta)$ зависит, очевидно, от значений ε и Θ. Подбираются такие значения ε и Θ, чтобы величина $\sum_{j=1}^{m} d(\gamma_{i_3}, \gamma(\varepsilon; \Theta))$ была минимальна.

Для наилучшего выбора вектора коэффициентов Θ можно использовать также так называемый «метод голосования». При любом $\varepsilon > 0$ с помощью линейной функции $\widehat{f}(X; \Theta) =$

---

[1] Алгоритм основан на результатах, изложенных в работе: *Киселев Н. И.* Экспертно-статистический метод определения функции предпочтения по результатам парных сравнений объектов // Алгоритмическое и программное обеспечение прикладного статистического анализа: Ученые записки по статистике. — М.: Наука, 1980. — Т. 36. — С. 111—122.

$= \sum_{l=1}^{p} \theta_l x^{(l)}$ строится разбиение n объектов следующим образом. Пусть в разбиениях классы занумерованы и $\gamma_{j_3}^{(\nu)}$ — ν-й класс в j-м экспертном разбиении. Для любого объекта X_i подсчитывается величина

$$\Gamma\left(X_i,\ \gamma_{j_3}^{(\nu)}\right) = \sum_{X_l \in \gamma_{j_3}^{(\nu)}} \tilde{\gamma}(X_i,\ X_l),$$

где

$$\tilde{\gamma}(X_i;\ X_l) = \begin{cases} 1, \text{ если } \left| \sum_{k=1}^{p} \theta_k \left(x_i^{(k)} - x_l^{(k)}\right) \right| \leqslant \varepsilon; \\ 0, \text{ если } \left| \sum_{k=1}^{p} \theta_k \left(x_i^{(k)} - x_l^{(k)}\right) \right| > \varepsilon. \end{cases}$$

Объект X_i относится к тому классу, для которого величина $\Gamma\left(X_i,\ \gamma_{j_3}^{(\nu)}\right)$ максимальна. Полученное разбиение обозначим через $\gamma_j\ (\varepsilon;\ \Theta)$. Параметры ε и Θ подбираются из условия минимизации величины $\sum_{j=1}^{m} d\left(\gamma_{j_3}, \gamma_j\ (\varepsilon;\ \Theta)\right)$ (при наличии априорных «весов компетентности» $v_1, \ldots, v_m$ минимизируется взвешенная сумма $\sum_{j=1}^{m} v_j \cdot d\left(\gamma_{j_3},\ \gamma_j\ (\varepsilon;\ \Theta)\right)$. Используется алгоритм эвристического типа.

Замечание. Выше отмечалось, что успех описываемого подхода целиком зависит от качества экспертной части исходной информации. Поэтому прежде чем непосредственно приступить к процедурам оценивания параметров Θ целевой функции, необходимо тщательно исследовать структуру и степень согласованности экспертных мнений. В варианте балльных оценок это сводится в основном к анализу резко выделяющихся наблюдений [11, § 11.5]. В варианте ранжировок используется аппарат ранговой корреляции [12, гл. 2] в первую очередь для того, чтобы проверить гипотезу об отсутствии какой бы то ни было согласованности в упорядочениях различных экспертов (см. также [9, с. 212—213]). В варианте парных сравнений исследуется структура попарных расстояний между экспертными разбиениями на классы.

Иногда удобно пользоваться *единым вариантом* экспертной оценки выходного качества объектов. В случае балльных оценок после исключения аномальных некомпетентных мнений пользуются средними арифметическими (усреднение по всем оставшимся экспертам) баллами для каждого объек-

та. В случае ранжировки и парных сравнений объектов следует пользоваться *медианными оценками*: каждому объекту приписывается ранг, равный медиане ряда рангов, присвоенных ему всеми экспертами; в качестве единого разбиения используется медианное разбиение, определяемое как решение оптимизационной задачи вида

$$\sum_{j=1}^{m} d(\gamma,\ \gamma_{j\text{э}}) \to \min_{\gamma}$$

[9, § 4, гл. 3].

15.4. Применение экспертно-статистического метода построения латентного интегрального показателя к решению практических задач

15.4.1. Построение целевой функции для оценки уровня мастерства спортсменов в игровых видах спорта (на примере «АИС-ХОККЕЙ-73»). Знание целевой функции позволяет в данном случае: 1) производить формализованную оценку мастерства хоккеиста, проявленного им в данном матче или в серии матчей, основанную только на знании отдельных числовых показателей, характеризующих его игру; 2) наиболее целесообразно строить индивидуальные планы тренировок, особое внимание уделяя совершенствованию тех компонентов игры, которые вошли в целевую функцию с относительно большими весами и за счет которых, следовательно, можно добиться наиболее существенного прироста в оценках мастерства.

Как и в любой работе такого профиля, в данной были последовательно реализованы следующие семь основных этапов:

этап 1: постановка задачи;

этап 2: предварительный отбор входных параметров;

этап 3: организация экспертных обследований;

этап 4: организация службы наблюдений, т. е. съема значений входных признаков;

этап 5: вывод целевой функции (определение ее общего вида и вычисление весовых коэффициентов);

этап 6: экспериментальная проверка адекватности целевой функции;

этап 7: рабочая эксплуатация целевой функции.

Ход и результаты данного исследования подробно описаны в [4].

15.4.2. Об использовании экспертно-статистического метода в анализе макроструктуры фондов потребления. В [173]

отражен опыт применения описанного выше экспертно-статистического метода (ЭСМ) для решения одной из задач Международного проекта «Plan/Cons» («Критерии выбора между рыночной и внерыночной формами удовлетворения потребностей населения»), осуществлявшегося в начале 70-х годов под эгидой ЮНЕСКО [1]. Обусловленные спецификой анализируемой задачи принципиальные трудности реализации ЭСМ (см. ниже) в данном случае существенно снизили практическую ценность его прикладного «выхода».

Смысл и место целевой функции в задаче оптимизации структуры потребления. В данном случае целевую функцию (функцию общественного благосостояния) следует интерпретировать не как неизменную во времени объективно существующую универсальную характеристику благосостояния общества, но лишь как удобный вспомогательный аппроксимирующий инструмент при решении задач оптимизации структуры потребления.

Все этапы применения описанного здесь формального аппарата должны сопровождаться проведением подробнейшего географического, политического, экономического, социологического, психологического и историко-этнографического анализа различных аспектов этой сложной комплексной проблемы (при отборе стран — объектов исследования; при отборе входных параметров; при выборе общего вида аппроксимации и т.д).

Требование однородности объектов по неучтенным переменным. Несмотря на то что вектор входных переменных $(x^{(1)}, \ldots, x^{(p)})$ должен отражать структуру потребления благ и услуг, понимаемых в самом широком смысле, ряд важных факторов и переменных остается при этом за рамками исследования К таким факторам относятся политические, географические, психологические, историко-этнографические и другие характеристики стран. Поэтому для того, чтобы предлагаемый метод был эффективным, он должен применяться лишь к совокупности стран, приблизительно однородных в отношении упомянутых выше неучтенных факторов. Во всяком случае бессмысленно было бы сопоставлять с помощью экспертно-статистической аппроксимации целевой функции страны различных формаций, скажем, социалистические и капиталистические.

---

[1] С. А. Айвазян входил в состав рабочей группы СССР и являлся автором предложенного этой группой варианта решения задачи моделирования, оценки и анализа тенденций структуры общественных фондов потребления стран.

О выборе входных параметров. При реализации экспертно-статистического метода в данной задаче встретились принципиальные трудности. При выборе входных параметров приходится одновременно считаться с двумя противоречивыми требованиями. С одной стороны, для достаточно полной характеристики структуры потребления, ее прогрессивности желательна весьма насыщенная система показателей, отражающих соотношение отдельных частей потребления на разных уровнях агрегации и при различных аспектах классификации.

С другой стороны, специфика данной задачи такова, что предоставляет исследователю крайне скудное количество исследуемых объектов (однородных стран), а потому вынуждает ограничиться лишь небольшим числом наиболее информативных агрегированных показателей. Невозможно обойти ограничение, в соответствии с которым число неизвестных параметров целевой функции не может превосходить числа обследуемых объектов, а это накладывает ограничение на размерность p вектора входных параметров.

Поэтому конкретная реализация экспертно-статистического метода построения целевой функции предусматривает необходимость максимального сжатия информации, содержащейся в достаточно развернутой системе показателей, и, в частности, переход к *небольшому* числу (не превышающему число анализируемых стран) *наиболее существенных* с точки зрения характера и степени прогрессивности структуры потребления показателей. В качестве таких показателей были предварительно выбраны:

$x^{(1)}$ — доля общественных фондов потребления — сумма поступлений населению, финансируемая из коллективных источников государственного бюджета, средств социального обеспечения, средств предприятий, организаций и профсоюзов, — и попадающих к населению в двух основных видах — товары и услуги и выплаты (трансферты). К первому виду поступлений относятся бесплатные или отпускаемые по льготным ценам товары и услуги, потребляемые населением в таких областях, как образование, здравоохранение, социальное обеспечение (учреждения для детей, престарелых и инвалидов), культура и информация, спорт, досуг и развлечения, жилище. Ко второму виду поступлений относятся выплаты престарелым и нетрудоспособным членам общества, учащимся: пенсии, всякого рода пособия и стипендии;

$x^{(2)}$ — удельный вес общественной формы организации потребления. Под общественной формой организации потребления подразумевается *коллективное* потребление това-

ров и услуг в таких областях, как образование, социальное обслуживание детей и престарелых в специальных учреждениях, здравоохранение, общественное питание, зрелищные мероприятия, общественный транспорт, культурно-просветительная работа;

$x^{(3)}$ — удельный вес (в общем фонде расширенного потребления) пособий, стипендий, дотаций, бесплатных услуг и т.п. в системе государственной помощи семье, включая детские ясли, сады, школу, среднее специальное и высшее образование;

$x^{(4)}$ — удельные расходы (в общем фонде расширенного потребления) на здравоохранение и культуру, включая информацию, спорт, досуг и развлечения.

Помимо трудностей, связанных с необходимостью совмещения жесткого ограничения по числу оцениваемых стран[1] со стремлением иметь достаточно полную систему входных переменных, существенной помехой в успешной реализации ЭСМ в данной задаче являлась практическая невозможность получения согласованных компетентных мнений в экспертной части исходной информации: и по компетентности, и по согласованности своих оценок (даваемых, как правило, лишь в форме парных сравнений) эксперты-«межстрановики» существенно уступали экспертам, участвующим в решении задачи, описанной в п. 15.4.1.

О понятии прогрессивности структуры потребления. Само понятие прогрессивности макроструктуры потребления весьма условно и относительно. В частности, оно может иметь содержательный смысл лишь в пределах сравнительно о д н о р о д н о й по общественному и политическому устройству, по географическому положению и, в какой-то мере по масштабам, человеческому и природному потенциалу группы стран и лишь в пределах относительно небольшого, порядка 10—12 лет, отрезка времени.

С учетом этих оговорок представляется все-таки возможным и целесообразным сформулировать некоторые из основных критерийных соображений, опираясь на которые эксперт сможет произвести требуемую ранжировку группы социалистических (или капиталистических) стран по степени прогрессивности структуры их фондов потребления и даже оценить это качество в баллах (исходя из 10-балльной системы оценок)

---

[1] Оцениваемые страны были разбиты (подобно хоккеистам, для которых строили целевую функцию отдельно по защитникам и нападающим) на две группы по семь объектов в каждой 1) НРБ, ВНР, ГДР, ПНР, СРР, СССР, ЧССР, 2) Бельгия, Франция, Финляндия, Италия, Норвегия, ФРГ, Швейцария.

Это, во-первых, анализ и оценка принятой в данной стране системы ценностей и общественных институтов и, в частности, принятая там трактовка и реализация таких понятий, как социальная справедливость, свободное и всестороннее развитие личности, психологическое равновесие индивидуума: различные положительные и отрицательные аспекты влияния окружающей среды на человека, урбанизация и т. п.

Во-вторых, это достигнутый к настоящему моменту уровень благосостояния населения и перспективы (на ближайшие 10—12 лет) его повышения.

Построенную в данной схеме с помощью ЭСМ целевую функцию предполагалось использовать при анализе, сопоставлении и частичном регулировании макроструктуры фондов потребления рассматриваемых стран.

15.4.3. Построение сводного показателя эффективности деятельности промышленного предприятия (см. также [87]). Объектами исследования являются 17 промышленных предприятий, специализирующихся на выпуске асинхронных двигателей переменного тока различного назначения. Цель исследования — построение единого сводного показателя экономической эффективности работы предприятия в форме линейной функции от ряда частных показателей экономической эффективности. С помощью сочетания экспертного анализа и статистических методов снижения размерности (метода экстремальной группировки признаков, метода главных компонент, см. гл. 13, 14) из априорного набора, состоящего из 22 частных показателей эффективности, было оставлено в качестве входных переменных экспертно-статистического метода восемь:

$x^{(1)}$ — удельный вес продукции высшей категории качества в общем объеме товарной продукции (ТП) предприятия;

$x^{(2)}$ — динамика[1] выпуска ТП на 1 рубль затрат;

$x^{(3)}$ — выработка нормативно-чистой продукции (НЧП) на единицу промышленно-производственного персонала (ППП);

$x^{(4)}$ — выполнение плана выпуска НЧП на единицу ППП;

$x^{(5)}$ — динамика фондоотдачи;

$x^{(6)}$ — выполнение плана выпуска ТП;

[1] Показатель динамики исчисляется как отношение прироста анализируемого показателя в данном периоде к его величине в предшествующем.

$x^{(7)}$ — выполнение плана по оборачиваемости нормируемых оборотных средств (отношение фактического числа оборотов к нормативному),

$x^{(8)}$ — выполнение плана по балансовой прибыли.

Из двенадцати привлеченных к задаче экспертов пять дали оценку интегральной эффективности деятельности анализируемых предприятий (по результатам их работы в 1982 и 1983 гг.) в десятибалльной системе, остальные семь проранжировали предприятия, причем четверо из них дополнительно представили результаты своих парных сравнений. Поэтому реализовывались все три версии оценивания неизвестных коэффициентов линейной целевой функции. Приведем здесь для примера один из полученных с помощью ЭСМ вариантов решения, а именно вариант, ориентированный на экспертные балльные оценки только одного из экспертов: $f(X; \Theta) = -2{,}94 + 0{,}10x^{(1)} + 0{,}07x^{(2)} + 0{,}55x^{(3)} - 0{,}13\ x^{(4)} + 0{,}06\ x^{(5)} - 0{,}03\ x^{(6)} - 0{,}07\ x^{(7)} - 0{,}02\ x^{(8)}$.

Мера согласованности балльного оценивания предприятий, произведенного с помощью этого эксперта и с помощью данной целевой функции, характеризуется величиной коэффициента корреляции, равной 0,77. Обращает на себя внимание тот факт, что формализация критерийных установок эксперта показала практическое игнорирование им при формировании интегральной оценки эффективности функционирования предприятия всех частных показателей выполнения плана (т. е. входных переменных $x^{(4)}$, $x^{(6)}$ $x^{(8)}$). В то же время если бы этому специалисту предложили непосредственно (экспертно) оценить значимость коэффициентов при этих переменных, то результат, можно не сомневаться, был бы совсем иным!

ВЫВОДЫ

1. Задача построения не поддающегося непосредственному измерению *интегрального* (агрегатного) сводного показателя y эффективности функционирования (качества) объекта по заданным значениям частных критериальных характеристик $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ анализируемого свойства может рассматриваться как задача снижения размерности исследуемого признакового пространства $\Pi^{(p)}(X)$ *до единицы*. Эта же задача может быть сформулирована в терминах построения целевой функции анализируемого обобщенного свойства исследуемых объектов.
2. *Целевой функцией* исследуемого обобщенного свойства объекта, характеризуемого значениями $x^{(1)}, \ldots, x^{(p)}$ его

частных критериальных характеристик, называется любое преобразование $\varphi(x^{(1)}, \ldots, x^{(p)})$, сохраняющее заданное соотношение порядка между анализируемыми объектами, т. е. обладающее тем свойством, что из $O_{i_1} \succ O_{i_2} \succ \ldots \succ O_{i_n}$ с необходимостью следует выполнение неравенств

$$\varphi(X_{i_1}) \geqslant \varphi(X_{i_2}) \geqslant \ldots \geqslant \varphi(X_{i_n})$$

и, наоборот (здесь знак «$\succ$» означает «не хуже», а X_j, как обычно, вектор $(x_j^{(1)}, \ldots, x_j^{(p)})'$).

3. Если рассмотреть *линейную модель* целевой функции, то задача ее определения (или, что то же, задача построения интегрального показателя y) сводится к определению весовых коэффициентов $\theta_1, \theta_2, \ldots, \theta_p$ в формуле $\varphi(X) = \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}$. Статистическая практика свидетельствует, что от экспертов гораздо проще получить информацию, относящуюся к сравнению объектов по анализируемому *интегральному свойству*, чем к сравнению удельных весов θ_j влияния на него отдельных частных критериальных показателей.
4. Базовая идея *экспертно-статистического метода* построения единого сводного показателя эффективности функционирования (качества) объекта заключается в «настройке» искомых коэффициентов θ_j целевой функции $\varphi(X)$ на заданную (в различной форме) экспертную информацию, относящуюся к сравнению статистически обследованных объектов по анализируемому интегральному свойству. Название метода объясняется тем, что его реализация основана как на *статистической информации* об объектах (это данные $X_1, X_2, \ldots, X_n$ о значениях их частных критериальных показателей), так и на *экспертной* (это представленные в той или иной форме экспертные оценки анализируемого интегрального свойства y).
5. Вычислительная реализация экспертно-статистического метода (т. е. алгоритм определения искомых «весов» θ_j) сводится к известному *методу наименьших квадратов* лишь в тех сравнительно редких случаях, когда от экспертов удается получить *балльные оценки* $y_{1э}, \ldots, y_{nэ}$ анализируемого интегрального свойства *по каждому из исследуемых объектов*. Если же в распоряжении исследователя лишь *сравнительные оценки* объектов по анализируемому свойству (упорядочения, парные сравнения, классификации), то вычислительная процедура по определению коэффициентов θ_j существенно усложняется

(ее описание и обоснование требуют специальной разработки).

6. Экспертно-статистический метод имеет широкий диапазон возможных применений, однако необходимым условием его достоверности и эффективности является *четкое определение анализируемого интегрального свойства и компетентность используемых экспертных мнений.*

Глава 16. МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ

Рассматриваются методы обработки таблиц экспериментальных данных, заданных в виде одной или нескольких матриц мер различия (или близости) между n объектами

$$\Delta^{(l)} = (\delta_{ij}^{(l)});\ i,\ j = 1,\ \ldots,\ n,\ l = 1,\ \ldots,\ k.$$

Таким образом, в общем случае имеем дело с трехвходовой матрицей, называемой еще «ящиком» экспериментальных данных.

Индекс l указывает, что l-я порция (матрица) близостей получена при проведении измерений на том же множестве объектов (единиц обследования, измерения), что и все другие матрицы $\Delta^{(r)}$ $(r \neq l)$, но с помощью некоторого l-го способа измерения, l-го эксперта, в l-й момент времени и т. д., т. е. l определяет условия, в которых проводилось формирование матрицы $\Delta^{(l)}$.

Основное внимание будем уделять случаю, когда удаленности (близости) δ_{ij} измерены либо в количественной шкале (интервальной или шкале отношений), либо в ординальной. В последнем случае для анализа играет роль только порядок расположения величины δ_{ij}. Цель методов, составляющих многомерное шкалирование (МШ), состоит в том, чтобы отобразить информацию о конфигурации точек, заданную матрицами близостей Δ, в виде геометрической конфигурации из n точек в многомерном пространстве. Если имеется несколько матриц $\Delta^{(l)}$, $l = 1,\ \ldots,\ k$, то одновременно строится и геометрическая конфигурация из k точек в пространстве той же размерности, каждая из которых является «образом» совокупности условий измерения (эксперта, измерительного прибора и т. д.). Такие методы носят название «многомерное шкалирование индивидуальных различий» (МШИР). Это отображение достигается путем приписывания каждому из объектов наблюдения (и, если $k > 1$, каждому из условий измерения) q-мерного вектора характеризующих его количественных показателей. Компоненты этих векторов определяются таким образом,

чтобы расстояния или близости (например, скалярные произведения) между точками (образами объектов) в пространстве отображения в среднем мало отличались от матриц $\Delta^{(l)}$ в смысле некоторого критерия. Размерность пространства q либо задается заранее, либо определяется в процессе решения задачи МШ или МШИР.

16.1. Метрическое многомерное шкалирование

16.1.1. Статистическая модель метрического МШ. В случае метрического МШ предполагается, что элементы единственной матрицы удаленностей Δ есть расстояния, измеренные с некоторой ошибкой, между объектами исследуемой совокупности, которые рассматриваются как точки в некотором q-мерном пространстве R^q:

$$\delta_{ij} = d_{ij} + \varepsilon_{ij}, \tag{16.1}$$

где d_{ij} — расстояние между точками X_i и X_j; ε_{ij} — ошибка измерения.

Обычно, хотя это и не обязательно, пространство R^q предполагается евклидовым, тогда $d_{ij} = \left(\sum_{r=1}^{q} (x_i^{(r)} - x_j^{(r)})^2\right)^{1/2}$. Дальше в данном параграфе будем иметь дело только с евклидовой метрикой.

16.1.2. Классическая модель и решение задачи метрического МШ. Описанные далее модель и способ определения координат точек $X_1, \ldots, X_n$ подробно рассмотрены в работах [318, 61, 152]. В данной модели предполагается, что ошибки измерения $\varepsilon_{ij} = 0$ $(i, j = \overline{1, n})$, так что δ_{ij} — это в точности евклидовы расстояния.

Метод определения координат точек $X_1, \ldots, X_n$ (с точностью до ортогонального вращения) и заодно размерности пространства, в которое они отображаются, основан, однако, не на непосредственном использовании матрицы Δ, а на преобразовании ее в матрицу $\mathbf{B}$ скалярных произведений центрированных векторов

$$b_{ij} = (X_i - \overline{X})'(X_j - \overline{X}). \tag{16.2}$$

Переход от матрицы исходной информации Δ к матрице $\mathbf{B}$ производится следующим образом. Оказывается

$$b_{ij} = \frac{1}{2}\left(-d_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n^2}\sum_{i,j} d_{ij}^2\right). \tag{16.3}$$

Процедура перехода от Δ к $\mathbf{B}$ называется *двойным центрированием* Δ. Матрица $\mathbf{B}$ размера $(n \times n)$ обладает следующими свойствами:

1) $\mathbf{B}$ неотрицательно определена;
2) ранг матрицы $\mathbf{B}$ равен размерности искомого пространства отображения;
3) ненулевые собственные числа матрицы $\mathbf{B}$, упорядоченные в порядке убывания, совпадают с соответствующими собственными числами матрицы $\mathbf{S} = \mathbf{X}\mathbf{X}'$, где $\mathbf{X}$ — центрированная матрица данных (неизвестная нам), т. е. матрица, элементы i-го столбца которой и являются координатами вектора X_i. Матрица $\mathbf{S}/n$ есть матрица ковариаций для X;
4) пусть U_r есть r-й собственный вектор матрицы $\mathbf{S}$, соответствующий r-му собственному числу λ_r, тогда вектор значений r-й главной компоненты будет $\mathbf{z}_r = \mathbf{X}' U_r$.

В то же время пусть $\mathbf{y}_r$ — r-й собственный вектор матрицы $\mathbf{B}$, соответствующий тому же самому собственному значению λ_r, т. е.

$$\mathbf{B}\mathbf{y}_r = \lambda_r \mathbf{y}_r, \tag{16.4}$$

тогда $\mathbf{z}_r = \sqrt{\lambda_r}\mathbf{y}_r$.

Из свойства 4) следует, что, решая проблему собственных чисел и собственных векторов для матрицы $\mathbf{B}$ и ограничиваясь ненулевыми собственными числами $\lambda_1, \ldots, \lambda_q$, получим координатное представление точек в пространстве главных компонент, основываясь на формуле (16.4);
величину размерности такого пространства, равную числу положительных собственных чисел матрицы q.

Элементы матрицы $\mathbf{B}$ могут быть представлены в виде

$$b_{ij} = \sum_{r=1}^{q} z_i^{(r)} z_j^{(r)}.$$

Очевидно, решение $\mathbf{Z}$ является линейной функцией $\mathbf{X}$ и определяется лишь *с точностью до ортогонального преобразования*, поскольку, применяя к матрице $\mathbf{Z}$ преобразование вращения, получим, что преобразованная матрица $\mathbf{Z}^*$ столь же точно восстанавливает матрицу $\mathbf{B}$, как и матрица $\mathbf{Z}$. Такое шкалирование можно назвать *линейным*.

16.1.3. Погрешность аппроксимации. Оптимальность линейного метрического МШ. Если возьмем число собственных векторов матрицы $q^* < q$, то получим некоторое приближение для элементов b_{ij}:

$$b_{ij}^* = \sum_{r=1}^{q^*} z_i^{(r)} z_j^{(r)}. \tag{16.4'}$$

Как следует из экстремальных свойств главных компонент (см. гл. 13),

$$\sum_{i,j}(b_{ij}-b^*_{ij})^2=\sum_{i=1}^{q-q^*}\lambda^2_{q^*+i}, \tag{16.5}$$

и это минимальное значение погрешности, которое может быть достигнуто при аппроксимации матрицы **B** матрицей **M** ранга q^*, т. е. матрицей, представимой в виде $\mathbf{M}=\sum_{i=1}^{q^*}c_i^2\mathbf{t}_i\mathbf{t}_i'$ (где $\mathbf{t}_i$ — n-мерные ортонормированные векторы), если измерять погрешность аппроксимации величиной

$$\tau^2(\mathbf{B},\ \mathbf{M})=\sum_{i,j}(b_{ij}-m_{ij})^2. \tag{16.6}$$

Заметим, что решение $m_{ij}=b^*{}_{ij}$, где $b^*{}_{ij}$ определены равенством (16.4), доставляет *глобальный минимум* критерию (16.6), хотя координатные векторы Z_1, ..., Z_{q^*} являются линейными функциями от X. Этот результат носит название теоремы Эккарта — Юнга [1].

На практике размерность пространства отображения q^* выбирают из тех же соображений, как и в анализе главных компонент, т. е. руководствуясь величиной объясненной доли следа.

16.1.4. Возможности расширения применимости линейного метрического МШ. Проблема аддитивной константы. Применение алгоритма линейного метрического шкалирования, строго говоря, будет корректным при выполнении следующих условий: все d_{ij} — евклидовы расстояния, и эти расстояния измерены без ошибки. Об устойчивости алгоритма к ошибкам свидетельствует значительное количество удачных его применений [90, 61, 89].

В случае, если различия δ_{ij} не являются евклидовыми расстояниями, матрица **B** может не быть положительно определенной. Прием, который используется в линейном метрическом МШ для преодоления этого, заключается в переходе к модели с так называемой аддитивной константой

$$\delta_{ij}+a\approx d_{ij}. \tag{16.7}$$

---

[1] *Eckart C., Young G.* The approximation of one matrix by another of lower rank // Psychometrika. — 1936. — Vol. 1. — P 211—318.

Очевидно, существует такое значение $a > 0$, что для величин d_{ij} будет выполняться неравенство треугольника, т.е. они будут расстояниями. В частности, это будет, если

$$\widehat{a} = (-1)\max_{h,\,i,\,j}(\delta_{hj} - \delta_{hi} - \delta_{ij}).$$

Значение $\widehat{a}$ есть минимальное значение константы a, при котором выполняется неравенство треугольника [201] для всех троек объектов из преобразованнои матрицы **D**.

Однако из выполнения неравенства треугольника еще не следует, что величины d_{ij} можно рассматривать как *евклидовы* расстояния и, следовательно, нельзя гарантировать неотрицательной неопределенности матрицы **B**, получаемой в результате процедуры двойного центрирования. Поэтому необходимо выбрать аддитивную константу таким образом, чтобы, с одной стороны, обеспечить положительную определенность матрицы **B** (или хотя бы небольшие значения модулей отрицательных собственных чисел), а с другой стороны, не увеличить существенно число значимых по величине положительных собственных чисел матрицы **B**, т. е размерность пространства отображения (с ростом a она будет расти). Подробнее о подходах к решению этой проблемы см [299]

Недостатком, ограничивающим практическое применение метода метрического МШ, является трудность работы с пропущенными данными, т. е. в случае, когда часть значений мер различия отсутствует. Тогда неясно, как корректно осуществить переход от матрицы **Δ** к **B** В то же время для нелинейного подхода к МШ и для неметрического МШ отсутствие части данных практически не сказывается на результатах.

16.1.5. Нелинейные методы метрического МШ. Эти методы основываются на получении матрицы путем прямой минимизации критерием вида

$$Q_1(\mathbf{Z}) = c_1(\Delta)\sum_{i\ j} v_{ij}(\delta_{ij} - d_{ij})^2 \qquad (16\ 8)$$

или

$$Q_2(\mathbf{Z}) = c_2(\Delta)\sum_{i\ j} v_{ij}(\delta_{ij}^2 - d_{ij}^2)^2. \qquad (16.8')$$

Семейство критериев вида (16 8) с различным выбором весов v_{ij} рассматривается в [300, 9, 152, 81] (см также § 13 6). Критерий вида (16 8′) предложен в [9, 329]. Вычислительные аспекты, связанные с минимизацией (16 8), описаны в

§ 13 6, некоторые другие подходы, например использование метода сопряженных градиентов, описаны в [152].

Веса v_{ij} в критерии (16.8) обычно выбирают в одной из следующих форм. $v_{ij}=1/\delta_{ij}$, $v_{ij}=\delta_{ij}$ (см также § 13.6). Вид критериев типа (16 8) аналогичен виду классического критерия неметрического шкалирования типа «стресс» (stress) [1]. Нормирующие константы подбираются так, чтобы, во первых, критерий стал однородным по δ_{ij}, и во-вторых, отражал некоторую относительную величину качества аппроксимации. Например, в критерии Сэммона (см § 13.5) вес $v_{ij}=1/\delta_{ij}$ и $c(\Delta)=1/\sum_{i,j}\delta_{ij}$. Наличие нормирующей константы не влияет, однако на получение минимизирующего решения, поскольку величины δ_{ij} считаются неизменными в процессе минимизации (в отличие от процедур неметрического МШ)

В качестве расстояний d_{ij} не обязательно брать евклидовы, можно использовать например, метрику Минковского [152]

Решение задачи шкалирования, полученное классическим методом, часто используется как начальная конфигурация для минимизации указанных критериев

При метрическом МШ, основанном на критериях типа (16 8), (16 8′), уже можно обрабатывать матрицы Δ с пропущенными элементами Для этого суммирование в (16.8) и (16 8′) достаточно проводить только для тех пар объектов, для которых удаленности измерены Экспериментально показано, что качество восстановления конфигурации будет почти таким же, как для полной матрицы, даже при достаточно большом числе пропусков (порядка 1/3 расстояний для каждого объекта) [90].

16.2. Неметрическое многомерное шкалирование [307, 261, 260, 152]

16.2.1. Структурная модель. В неметрическом МШ предполагается, что различия (близости) δ_{ij} измерены в ординальной шкале, так что важен только ранговый порядок различий, а сами их численные значения не так важны. Процедуры неметрического МШ стремятся построить такую геометрическую конфигурацию точек в q-мерном пространстве, чтобы ранговый порядок попарных расстояний, между ними

[1] Основное отличие критериев типа (16 8) (16 8′) от критериев типа «стресс» (см (16 10), (16 10′)) нормирующие множители $c_i(\Delta)$ зависят от искомых координат точек и меняются в процессе работы алгоритма оптимизации

совпадал по возможности с ранговым порядком различий, т. е отобразить неметрическую (ранговую) информацию в метрической шкале. Поскольку ранговый порядок не меняется при любом монотонно возрастающем преобразовании, задаваемом функцией $f(\cdot)$, то приходим к следующей структурной модели. $f(\delta_{ij}) \approx d_{ij}$. Это означает, что процедура построения подходящей геометрической конфигурации включает в себя не только подгонку координат точек-образов, но и самой функции $f(\cdot)$.

Дальше через $\widehat{d}_{ij}$ будем обозначать значение $f(\delta_{ij})$, т. е. $\widehat{d}_{ij} = f(\delta_{ij})$. Для измерения того, насколько в среднем близки эти значения к аппроксимирующим их расстояниям d_{ij}, используются различные критерии, например

$$S_1(\mathbf{Z}, f) = \sqrt{\sum_{i,j} (\widehat{d}_{ij} - d_{ij})^2} \Big/ \sqrt{\sum_{i,j} d_{ij}^2}. \qquad (16.9)$$

Это известный «стресс»- критерий (форма 1), предложенный Краскалом [261]. Используются и его модификации [329, 89]:

$$S_2(\mathbf{Z}, f) = \sqrt{\sum_{i,j} (\widehat{d}_{ij} - d_{ij})^2} \Big/ \sqrt{\Sigma (d_{ij} - \bar{d})^2}; \qquad (16.10)$$

$$S_3(\mathbf{Z}, f) = \sqrt{\sum_{i,j} \frac{(\widehat{d}_{ij} - d_{ij})^2}{d_{ij}^2}}, \qquad (16.10')$$

где $\bar{d} = \frac{1}{n^2} \sum_{i,j} d_{ij}$ —среднее значение расстояния.

Функция $f(\cdot)$ может задаваться как в параметрическом виде, так и непараметрически Для последнего случая Краскалом предложен метод получения геометрической конфигурации по критерию (16 9) или (16 10), который носит название шкалирования на основе монотонной регрессии [260].

При параметрическом задании функцию $f(\cdot)$ выбирают из некоторого параметрического семейства монотонных функций $f(\delta, \Theta)$ н, кроме q-мерных наборов координат, определяются и значения параметров Аддитивная константа, рассмотренная в п. 16 1.4 , может служить простейшим примером задания функции $f(\delta, \Theta)$. Эта константа является единственным оцениваемым параметром при таком заданнн функции. Большие возможиостн дает нспользование лннейной функцни

$$\widehat{d}_{ij} = a\delta_{ij} + b_i. \qquad (16.11)$$

Здесь уже имеется двумерный вектор параметров $\theta_1 = a$, $\theta_2 = b$

16.2.2. Некоторые замечания о вычислительной процедуре. Когда функция $f(\delta, \Theta)$ задана, критерий (16.9) (равно как и критерии (16.10), (16.10′)) можно переписать в виде $S_1(\mathbf{Z}, \Theta)$. Будем считать, что имеется единственное значение $\Theta_0 = \Theta_0(\mathbf{Z})$, которое минимизирует (16.9) при фиксированном $\mathbf{Z}$ (для функции вида (16.11) это, очевидно, имеет место).

Теперь, подставляя в $S_1(\mathbf{Z}, \Theta)$ минимизирующее значение Θ_0, видим, что для получения $\mathbf{Z}$ нужно минимизировать критерий

$$\tilde{S}(\mathbf{Z}) = S_1(\mathbf{Z}, \Theta(\mathbf{Z})). \tag{16.12}$$

Вычислим теперь градиент (16.12) по $\mathbf{Z}$. Имеем

$$\nabla\tilde{S}(\mathbf{Z}) = \frac{\partial S_1}{\partial \mathbf{Z}}(\mathbf{Z}, \Theta) + \frac{\partial S_1}{\partial \Theta}\Big|_{\mathbf{Z}, \Theta_0} \frac{\partial \Theta_0}{\partial \mathbf{Z}}.$$

Но, поскольку значение Θ_0 получено само из условия минимума $S_1(\mathbf{Z}, \Theta)$ по Θ, необходимо должно выполняться условие

$$\frac{\partial S_1}{\partial \Theta}\Big|_{\mathbf{Z}, \Theta_0} = 0$$

и, следовательно, выражение для градиента упрощается:

$$\nabla\tilde{S}(\mathbf{Z}) = \frac{\partial S_1}{\partial \mathbf{Z}}(\mathbf{Z}, \Theta_0). \tag{16.13}$$

Таким образом, каждая итерация в задаче минимизации критерия (16.10) может быть разбита на две фазы:

1) минимизация по Θ при заданном $\mathbf{Z}$. В случае функции (16.11) эта задача сводится к оценке a и b по методу наименьших квадратов и решается просто и однозначно,

2) минимизация $\tilde{S}(\mathbf{Z})$ при фиксированном Θ. Здесь, как правило, используется градиентная процедура. Затем происходит возврат к фазе 1. Весьма важным моментом на фазе 2 является выбор шага. По этому поводу и относительно других деталей вычислительной процедуры см. работу [89].

16.3. Шкалирование индивидуальных различий (ШИР)

В этом случае имеется k ($k > 1$) таблиц удаленностей или расстояний (если используется метрическое шкалирование). Будем обозначать различия между i-м и j-м объектами для

l-й таблицы ($l = \overline{1, k}$) через $\delta_{ij}^{(l)}$. В случае матриц расстояний структуриая модель (ШИР) предполагает, что расстояния $\delta_{ij}^{(l)}$ между точками для l-й матрицы могут быть представлены в виде взвешенного евклидова расстояния

$$\delta_{ij}^{(l)} \approx d_{ij}^{(l)} = \sqrt{\sum_{r=1}^{q} v_{lr} \left(z_i^{(r)} - z_j^{(r)}\right)^2}. \qquad (16.14)$$

В неметрическом случае структурная модель будет

$$f\left(\delta_{ij}^{(l)}\right) \approx d_{ij}^{(l)} = \sqrt{\sum_{r=1}^{q} v_{lr} \left(z_i^{(r)} - z_j^{(r)}\right)^2}. \qquad (16.15)$$

В метрическом случае и в предположении об отсутствии ошибок можно обобщить подход, рассмотрениый в § 16.1. Именно процедура двойного центрироваиия применяется для каждой из k матриц, что дает в результате набор уравнений $b_{ijl} \approx \sum_{r=1}^{q} z_i^{(r)} z_j^{(r)} v_{lr}$, где векторы Z_i центрированы.

Значения **Z** и **V** (**V** — матрица значений весов размером $k \times q$) получаются из минимизации, например, следующей функции потерь

$$S(\mathbf{Z}, \mathbf{V}) = \sum_{i, j, l} \left(b_{ijl} - \sum_{r=1}^{q} x_i^{(r)} x_j^{(r)} v_{lr} \right)^2. \qquad (16.16)$$

Вычислительные процедуры для ШИР приведены, например, в работах [317, 329, 152].

ВЫВОДЫ

1 *Многомерное шкалирование* — совокупность методов, позволяющих по заданной информации о мерах различия (близости) между объектами рассматриваемой совокупности приписывать каждому из этих объектов вектор характеризующих его количественных показателей; при этом размерность искомого координатного пространства задается заранее, а «погружение» в него анализируемых объектов производится таким образом, чтобы структура взаимных различий (близостей) между ними, измеренных с помощью приписываемых им вспомогательных координат, в среднем наименее отличалась бы от заданной в смысле того или иного функционала качества. Процеду-

ры многомерного шкалирования применяются, когда данные заданы в виде матрицы попарных расстояний между объектами или удаленностей или их порядковых отношений. В первом случае используются методы так называемого метрического шкалирования, а во втором — неметрического шкалирования.

2. Важной целью методов шкалирования — дать наглядное визуальное отображение данных в виде некоторой геометрической конфигурации точек.
3. Решения как в метрическом, так и в неметрическом случае неоднозначны — они определяются с точностью до поворота и переноса начала координат.
4. При наличии нескольких матриц расстояний (удаленностей), порядковых отношений этих удаленностей, задача шкалирования носит название задачи *шкалирования индивидуальных различий*. При этом, кроме образов объектов как точек в пространстве низкой размерности, можно получить и точки-образы для условий, породивших различные матрицы.
5. Вычислительные процедуры как в метрическом, так и в неметрическом случае весьма трудоемки (порядок числа умножений растет как n^2).

Глава 17. СРЕДСТВА АНАЛИЗА И ВИЗУАЛИЗАЦИИ НЕКОЛИЧЕСТВЕННЫХ ДАННЫХ

В данной главе рассматривается подход к анализу неколичественных данных, основанный на использовании методов анализа соответствий и оцифровки.

Анализ соответствий (АС) был введен и довольно широко используется в практическом анализе данных начиная с начала 60-х годов группой французских статистиков [191, 263]. Многие результаты, теоретически эквивалентные результатам АС, в особенности относящиеся к анализу двумерных таблиц сопряженности, неоднократно переоткрывались начиная с 30-х годов различными исследователями под названиями «дуальное шкалирование», «оптимальная оцифровка», «одиовременная линейная регрессия» и т.д. (см. библиографию в 12, гл. 3).

Несомненной заслугой французских статистиков является, помимо распространения АС на случай более чем двух переменных, широкое использование возможностей визуализации данных, предоставляемых АС.

В этой главе рассматривается применение АС для анализа двумерных частотных таблиц сопряженностей, т. е. собственно «классический» АС, введенный в [191]; распространение АС на анализ некоторых типов матриц данных с неотрицательными элементами; множественный анализ соответствий (МАС), т. е. методы АС в случае многомерных ($p > 2$) матриц данных с категоризованными переменными; методы оцифровки, отличные от МАС.

Как в АС, так и в МАС имеются определенные возможности включать, использовать и непрерывные переменные.

Рассмотрение АС для двухвходовых таблиц сопряженности, т. е. собственно АС, ведется здесь в основном. следуя стилю работ французских авторов (см., например, [263]). МАС вводится как некоторое обобщение метода главных компонент, что позволяет сразу же дать статистическую интерпретацию МАС.

17.1. Анализ соответствий для двухвходовых таблиц сопряженностей

17.1.1. Основные понятия анализа соответствий. Рассмотрим основные понятия АС: таблицы сопряженностей, профили, веса их, метрики.

Таблица сопряженностей. Пусть имеем в качестве объекта статистического анализа *двухвходовую таблицу сопряженностей* (ТС) (кросс-классификации) для двух категоризованных переменных x_1 и x_2 с i_1 и i_2 категориями соответственно. Эта таблица представляет собой матрицу $\mathbf{F}$ с l_1 строками и l_2 столбцами. Значением элемента (клетки) f_{ij} является вероятность одновременного наблюдения i-й категории признака x_1 и j-й категории признака x_2. Таким образом, с помощью этой таблицы полностью описывается совместное распределение двух категоризованных переменных x_1 и x_2.

На практике обычно приходится иметь дело с некоторой оценкой ТС, а именно с матрицей $\widehat{\mathbf{F}}$, элементы которой $\widehat{f}_{ij}$ представляют собой оценки соответствующих вероятностей $\widehat{f}_{ij}$ по выборке объема n, например, с помощью относительных частот $\widehat{f}_{ij} = n_{ij}/n$, где n_{ij} — частота появления события $x_1 = i$ и $x_2 = j$ (т. е. количество объектов с подобным сочетанием категорий) в выборке. Однако там, где это не связано с изучением выборочных свойств ТС, будем применять обозначения $\mathbf{F}$, f_{ij} и т. д.

В дальнейшем будем иногда использовать и частотную ТС, т. е. матрицу $\mathbf{N} = (n_{ij})$ $(i = \overline{1, i_1}, j = \overline{1, i_2})$, значениями элементов которой являются сами наблюдаемые частоты. Очевидно, что $\sum_{ij} n_{ij} = n$.

Анализу ТС $\widehat{\mathbf{F}}$ и $\mathbf{N}$ посвящено большое количество работ (см., например, [12, 21]; в этих же работах приведена и обширная библиография). Основная направленность обычного анализа ТС — проверить с помощью статистических критериев гипотезу о независимости переменных x_1 и x_2, и если они оказываются зависимыми, измерить с помощью какого-либо коэффициента связи степень их связи.

Методы АС применимы к ТС не только типа кросс-классификационных таблиц, но и таблиц $\mathbf{F}$ более общего вида, элемент f_{ij} которых можно рассматривать как степень связи, влияния строки i на столбец j или наоборот. Например, в качестве строк могут выступать страны мира, а в качестве столбцов — продукты питания, тогда элемент f_{ij} определяет долю j-го продукта питания в структуре питания жителей i-й страны. Другим важным примером является таблица — матрица межотраслевого баланса.

Профили. АС используется для объяснения структуры связей (соответствия) между категориями переменных x_1 и x_2. При этом категории рассматриваются как точки в некотором многомерном пространстве. Приведем теперь некоторые определения.

Профилем i-й строки ТС называется строка с элементами

$$\widehat{p}_{ij} = n_{ij}/n_{i\cdot}, \tag{17.1}$$

где

$$n_{i\cdot} = \sum_{j=1}^{i_2} n_{ij}. \tag{17.1'}$$

Очевидно, что $\widehat{p}_{ij}$ можно выразить и через элементы ТС относительных частот $\widehat{\mathbf{F}}$:

$$\widehat{p}_{ij} = \widehat{f}_{ij}/\widehat{f}_{i\cdot}, \tag{17.2}$$

где

$$\widehat{f}_{i\cdot} = \sum_{j=1}^{i_2} \widehat{f}_{ij}. \tag{17.2'}$$

Категорию i признака x_1 можно рассматривать как точку в пространстве R^{l_2} с компонентами $\widehat{p}_{i1}, \ldots, \widehat{p}_{il_2}$. Очевидно, при этом имеется одна связь между компонентами этой точки: $\sum_{j=1}^{l_2} \widehat{p}_{ij} = 1$. Аналогично можно ввести профили столбцов:

$$\widehat{q}_{ij} = n_{ij}/n_{\cdot j} = \widehat{f}_{ij}/\widehat{f}_{\cdot j}, \tag{17.3}$$

где

$$n_{\cdot j} = \sum_{i=1}^{l_1} n_{ij}, \quad \widehat{f}_{\cdot j} = \sum_{i=1}^{l_1} \widehat{f}_{ij}. \tag{17.3'}$$

Соответственно категории признака x_2 будем рассматривать как точки в l_1-мерном пространстве R^{l_1}, координаты которых задаются профилями (17.3). Очевидно, что $\sum_{i=1}^{l_1} \widehat{q}_{ij} = 1$ для всех $(j = \overline{1, l_2})$.

Вероятностный смысл профиля для категории i признака x_1, т. е. вектора с компонентами $\widehat{p}_{i1}, \ldots, \widehat{p}_{il_2}$, следует из того, что компонента $\widehat{p}_{ij}$ есть оценка условной вероятности для признака x_2 принять категорию j, если признак x_1 принял категорию i. Таким образом, это строка условных вероятностей (или их оценок).

Метрика χ^2. Для дальнейшего анализа категорий как точек в пространствах R^{l_1} и R^{l_2} необходимо ввести некоторую функцию расстояния между ними, т. е. метрику.

В АС используется χ^2-*метрика*. Расстояния между категориями признаков x_1 и x_2 в этой метрике задаются соответственно следующим образом:

$$d^2(i, i') = \sum_{j=1}^{l_2} \frac{1}{\widehat{f}_{\cdot j}} (\widehat{p}_{ij} - \widehat{p}_{i'j})^2; \tag{17.4}$$

$$d^2(j, j') = \sum_{i=1}^{l_1} \frac{1}{\widehat{f}_{i\cdot}} (\widehat{q}_{ij} - \widehat{q}_{ij'})^2. \tag{17.5}$$

Таким образом, метрика (17.4) есть просто взвешенная евклидова метрика в пространстве профилей строк R^{l_2} с весами, обратными относительной частоте категорий признака x_2. То же самое верно для метрики (17.5) (с заменой столбцов на строки и признака x_2 на x_1).

Одна из основных причин использования χ^2-метрики связана с тем, что она удовлетворяет свойству *инвариантности*

по отношению к слиянию строк (столбцов) с *одинаковыми* профилями, которое может быть сформулировано следующим образом:

а) пусть две строки i и i' (т. е. две категории признака x_1) имеют одинаковые профили; тогда, если объединить эти две категории в одну новую категорию i_0, расстояния между категориями признака x_2 не изменятся;

б) аналогично, если имеем два столбца j и j' с одинаковыми профилями и объединим категории j и j' в одну новую категорию j_0 (т. е. перейдем к новой ТС с $l_2 - 1$ категориями для признака x_2), то расстояния между строками, задаваемые формулой (17.4), не изменятся. Доказательство этого свойства несложно (см., например, [263]).

Веса профилей. Каждой из l_2 точек в пространстве R^{l_1} (т. е. профилям-столбцам) поставим в соответствие ее *вес* $f_{\cdot j}$ $(j = \overline{1, l_2})$, аналогично каждой из l_1 точек в пространстве R^{l_2} (т. е. профилям-строкам) поставим в соответствие вес $f_{i\cdot}$ $(i = \overline{1, l_1})$.

Итак, в результате имеем два взвешенных множества точек: одно — в пространстве R^{l_1} и другое — в пространстве R^{l_2}, расстояния между которыми задаются с помощью метрики χ^2 ((17.4), (17.5)). Суммарное представление введенных понятий дано в табл. 17.1.

Таблица 17.1

Пространство строк (R^{l_2})	Пространство столбцов (R^{l_1})
Количество точек l_1 Координаты точек — строки матрицы $\mathbf{F}_1 = \mathbf{D}_{l_1}^{-1}\mathbf{F}$ (профили строк матрицы $\mathbf{F}$), где $\mathbf{D}_{l_1} = \mathrm{diag}\,(f_{1\cdot}, \ldots, f_{l_1\cdot})$ Метрика (скалярное произведение, расстояние, норма) определяется матрицей $\mathbf{D}_{l_2}^{-1}$ Пусть $Z \in R^{l_2}$ и $U \in R^{l_2}$ Тогда $\|Z\|^2 = Z'\,\mathbf{D}_{l_2}^{-1}\,Z$ $d^2(Z, U) = (Z-U)'\,\mathbf{D}_{l_2}^{-1}\,(Z-U)$ Скалярное произведение векторов Z и U определяется как $(Z'\,\mathbf{D}_{l_2}^{-1}\,U)$ Веса точек — диагональные элементы матрицы $\mathbf{D}_{l_1}$	Количество точек l_2 Координаты точек — строки матрицы $\mathbf{F}_2 = \mathbf{F}'\,\mathbf{D}_{l_2}^{-1}$ (профили столбцов матрицы $\mathbf{F}$), где $\mathbf{D}_{l_2} = \mathrm{diag}\,(f_{\cdot 1}, \ldots, f_{\cdot l_2})$ Метрика (скалярное произведение, расстояние, норма) определяется матрицей $\mathbf{D}_{l_1}^{-1}$ Пусть $V \in R^{l_1}$ и $Y \in R^{l_1}$ Тогда $\|Y\|^2 = Y'\,\mathbf{D}_{l_1}^{-1}\,Y$ $d^2(V, Y) = (V-Y)'\,\mathbf{D}_{l_1}^{-1}\,(V-Y)$ Скалярное произведение векторов V и Y определяется как $(V'\,\mathbf{D}_{l_1}^{-1}\,Y)$ Веса точек — диагональные элементы матрицы $\mathbf{D}_{l_2}$

17.1.2. Проекции строк и столбцов. Связь с анализом главных компонент. Рассматривая профили строк и столбцов как точки в соответствующих пространствах R^{l_1} и R^{l_2}, дальше можно действовать несколькими способами, которые приводят к одинаковому результату.

Прежде всего для упрощения дальнейших выкладок нормируем профили строк (столбцов) так, чтобы χ^2-метрика стала обычной евклидовой (дальше $\mathbf{D}$, $\mathbf{D}_{l_i}$):

$$\begin{cases} \tilde{\mathbf{p}}_i = \mathbf{D}_2^{-1/2}\,\mathbf{p}_i \quad (i = \overline{1, l_1}); \\ \tilde{\mathbf{q}}_j = \mathbf{D}_1^{-1/2}\,\mathbf{q}_j \quad (j = \overline{1, l_2}). \end{cases} \tag{17.6}$$

Легко проверить, что евклидово расстояние между нормированными профилями строк (столбцов) совпадает с χ^2-расстоянием между соответствующими исходными профилями. Нормированные профили-строки являются строками матрицы $\mathbf{F}_r = \mathbf{D}_1^{-1}\mathbf{F}\mathbf{D}_2^{-1/2}$.

Введем теперь матрицу рассеивания $\mathbf{T}_r$ для нормированных профилей строк с учетом их весов

$$\mathbf{T}_r = \sum_{i=1}^{l_1} w_{ri}\,\tilde{\mathbf{p}}_i\,\tilde{\mathbf{p}}_i' = \sum_{i=1}^{l_1} \sqrt{w_{ri}}\,\tilde{\mathbf{p}}_i \sqrt{w_{ri}}\,\tilde{\mathbf{p}}_i'. \tag{17.7}$$

Матрица $\mathbf{T}_r$ имеет размеры $l_2 \times l_2$. Это аналог ковариационной матрицы системы из l_1 точек, но рассеивание измеряется не относительно их центра тяжести, а относительно нулевой точки. Будем теперь искать одномерную проекцию с вектором U, для которой рассеивание (дисперсия) образов точек максимально. Но это задача анализа главных компонент (см. гл. 13). В вычислительном отношении это приводит к решению проблемы собственных значений и векторов:

$$\mathbf{T}_r U = \lambda U. \tag{17.8}$$

С учетом того, что веса w_{ri} равны диагональным элементам матрицы $\mathbf{D}_1$, матрица $\mathbf{T}_r$ может быть представлена в виде

$$\mathbf{T}_r = \mathbf{F}_r'\mathbf{W}\mathbf{F}_r = \mathbf{D}_2^{-1/2}\mathbf{F}'\,\mathbf{D}_1^{-1}\mathbf{F}\mathbf{D}_2^{-1/2}. \tag{17.9}$$

Аналогично матрица рассеивания для нормированных профилей столбцов есть

$$\mathbf{T}_c = \mathbf{D}_1^{-1/2}\mathbf{F}\mathbf{D}_2^{-1}\mathbf{F}'\,\mathbf{D}_1^{-1/2}. \tag{17.9'}$$

Введем в рассмотрение матрицу

$$\mathbf{\Phi} = \mathbf{D}_1^{-1/2}\mathbf{F}\mathbf{D}_2^{-1/2}. \tag{17.10}$$

Тогда имеем $\mathbf{T}_r = \mathbf{\Phi}'\mathbf{\Phi}$ и $\mathbf{T}_c = \mathbf{\Phi}\mathbf{\Phi}'$.

Следовательно, матрицы $\mathbf{T}_r$ и $\mathbf{T}_c$ имеют одни и те же положительные собственные числа $\lambda_0 \geqslant \lambda_1 \geqslant \ldots \geqslant \lambda_{l^+} > 0$ и количество ненулевых собственных чисел $l^+ \leqslant \min(l_1, l_2)$. Собственные векторы матриц $\mathbf{T}_r$ и $\mathbf{T}_c$ с единичной нормой, соответствующие одному и тому же собственному значению $\lambda_k > 0$, связаны соотношением

$$\begin{cases} V_k = 1/\sqrt{\lambda_k}\, \mathbf{\Phi} U_k; \\ U_k = 1/\sqrt{\lambda_k}\, \mathbf{\Phi}' V_k. \end{cases} \tag{17.11}$$

При практических вычислениях, естественно, выбирается задача на собственные значения с минимальной размерностью

17.1.3. **Интерпретация главных компонент в анализе соответствий.** Легко видеть, что фактор $\Psi_k = \mathbf{D}_1^{-1/2} V_k$ пропорционален l_1-мерному вектору координат проекций нормированных профилей-строк $\mathbf{p}_i$ $(i = \overline{1, l_1})$ на вектор U_k. Аналогичный смысл, но для нормированных вектор-столбцов имеет вектор $\varphi_k = \mathbf{D}_2^{-1/2} U_k$. Действительно, вектор координат проекции нормированных профилей строк будет $\mathbf{F}_r U_k$. Несложные преобразования дают $\mathbf{F}_r U_k = \mathbf{D}_1^{-1} \mathbf{F} \mathbf{D}_2^{-1/2} U_k = \mathbf{D}^{-1/2} \mathbf{\Phi} U_k = 1/\sqrt{\lambda_k} \psi_k$. Но это значит, что координаты проекций точек из R^{l_2} (R^{l_1}) на направление, задаваемое собственным вектором U_k (V_k), пропорциональны (с множителем $1/\sqrt{\lambda_k}$) компонентам фактора в другом пространстве (R^{l_1} или R^{l_2}), соответствующем тому же самому собственному числу. Итак, координаты проекций получаются умножением факторов φ_k, ψ_k на $\sqrt{\lambda_k}$:

$\widetilde{\varphi}_k = \sqrt{\lambda_k}\, \varphi_k$ — координаты проекций из R^{l_1};

$\widetilde{\psi}_k = \sqrt{\lambda_k}\, \psi_k$ — координаты проекций из R^{l_2}.

Теперь, используя соотношения (17.11), имеем:

$$\begin{aligned} \widetilde{\psi}_{ki} &= 1/\sqrt{\lambda_k} \sum_{j=1}^{l_2} \frac{f_{ij}}{f_i} \widetilde{\varphi}_{kj} \quad (i = \overline{1, l_1}); \\ \widetilde{\varphi}_{kj} &= 1/\sqrt{\lambda_k} \sum_{i=1}^{l_1} \frac{f_{ij}}{f_{\cdot j}} \widetilde{\psi}_{ki} \quad (j = \overline{1, l_2}). \end{aligned} \tag{17.12}$$

Соотношения (17.11) можно интерпретировать следующим образом: проекция i-го профиля строки на ось U_k (равная $\widetilde{\psi}_{ki}$) с точностью до множителя $1/\sqrt{\lambda_k}$ (одинакового для всех $i = \overline{1, l_1}$) является взвешенным центром тяжести для

проекции профилей столбцов на ось V_k, вычисленным с весами $f_{ij}/f_{i\cdot}$, т. е. j-й вес равен j-й компоненте профиля i-й строки. Это свойство полностью характеризует факторы φ_k и ψ_k и может быть взято как исходное при определении АС.

Отметим теперь еще три свойства решений уравнений (17.11) и (17.8):

1) существуют факторы φ_0, ψ_0 с единичными компонентами и собственным числом $\lambda_0 = 1$ (*тривиальный фактор*). Это решение появляется в силу того, что сумма элементов для любого из профилей равна 1. Если при выводе уравнений для векторов U и V использовать разброс относительно центра тяжести, а не относительно нулевого профиля, как было сделано, то тривиальное решение не появилось бы;

2) все собственные числа $0 \leqslant \lambda_k \leqslant 1$. Количество ненулевых собственных чисел l^+, включая тривиальное, не превышает $l^+ = \min(l_1, l_2)$. Для суммы ненулевых собственных чисел имеет место равенство

$$\sum_{k=1}^{l^+} \lambda_k = X^2/n,$$

где X^2 — статистика χ^2, вычисленная для таблицы сопряженностей N,

3) имеет место следующее разложение матрицы $\mathbf{F}$ по системе факторов φ_k и ψ_k (см. [263, 12]):

$$f_{ij} = f_{i\cdot} f_{\cdot j}\left(1 + \sum_{k=1}^{l^+} \sqrt{\lambda_k}\, \psi_{ik}\, \varphi_{jk}\right). \qquad (17.13)$$

17.1.4. Присвоение числовых меток строкам и столбцам. Компоненты факторов φ_k, ψ_k можно рассматривать как наборы числовых меток, которые присваиваем строкам (столбцам) матрицы $\mathbf{F}$. Таким образом, можно говорить о квантификации (или оцифровке) строк (столбцов) матрицы $\mathbf{F}$. В случае, когда матрица $\mathbf{F}$ есть ТС, с помощью АС получается переход от неколичественных шкал для переменных x_1, x_2 к нескольким наборам количественных.

Заметим, что из соотношения (17.13) следует, что каждый из наборов меток φ_k, ψ_k обладает свойством наилучшего в смысле среднеквадратической ошибки взаимного прогноза. Действительно, i-я компонента φ_{ik} $(i = \overline{1, l_2})$ фактора φ_k пропорциональна условному математическому ожиданию фактора ψ_k при фиксировании i-й категории признака x_2. Аналогичным свойством обладают компоненты фактора ψ_k. Но условное математическое ожидание как раз и обладает

свойством наилучшего прогноза [16, 7, 12]. При этом уравнения регрессии φ_k по ψ_k и ψ_k по φ_k будут линейными.

Использование факторов φ_k и ψ_k. Как же предлагается использовать получаемые факторы φ_k, ψ_k ($k=\overline{1, l^+}$) в АС? Во-первых, их можно использовать, в силу (17.16), для аппроксимации элементов матрицы $\mathbf{F}$ (или $\widehat{\mathbf{F}}$, тогда над всеми величинами в (17.13) следует поставить символ $\frown$).

Во-вторых, и это основное использование получаемых факторов в АС, их используют для визуального отображения строк и столбцов на прямую или на плоскость. Для отображения на плоскость вычисляются факторы ψ_1, ψ_2 и φ_1, φ_2, соответствующие наибольшим собственным числам λ_1 и λ_2. Пара чисел $(\sqrt{\lambda_1}\,\psi_{i1},\ \sqrt{\lambda_2}\psi_{i2})$ $(i=1, l_1)$ служит координатами для i-и строки (i-й категории признака x_1). Соответственно пара чисел $(\sqrt{\lambda_1}\,\varphi_{j1},\ \sqrt{\lambda_2}\varphi_{j2})$ служит координатами для j-го столбца Далее проводится визуальный анализ получаемых конфигураций точек, соответствующих строкам и столбцам для выявления различных особенностей: наличие кластеров, скоплений некоторых точек-столбцов вблизи тех или иных строк, и наоборот.

17.2. Множественный анализ соответствий (МАС)

МАС является обобщением обычного АС на случай нескольких переменных, что можно сделать несколькими способами, которые приводят к эквивалентному результату. В случае $p = 2$ в любом случае придем к обычному АС [263].

Рассмотрим два эквивалентных подхода, ведущих к МАС Первый позволяет легко ввести расстояния между объектами и между категориями, второй рассматривает МАС как обобщение метода главных компонент и допускает прозрачную статистическую интерпретацию МАС. Другие возможные подходы к обобщению АС рассмотрены, например, в [263, 110]

17.2.1. Бинарная форма матрицы данных. Предположим, что исходные данные представлены в виде матрицы данных $\mathbf{X}$ и что все переменные, входящие в матрицу данных, являются категоризованными (или некоторые из них могут быть получены квантованием количественных непрерывных переменных). Представим все переменные в бинарной форме, т.е. переменной $x^{(i)}$ с числом категорий l_i поставим в соответствие набор из l_i бинарных переменных y_j^i $(j=\overline{1, l_i})$, таких,

что $y_j^i = 1$, если значение $x^{(i)}$ есть j-я категория и $y_j^i = 0$ — в противном случае Матрица данных в бинарной форме представляет собой матрицу $\mathbf{Y}$ размера $n \times m$, значениями элементов которой могут быть только 0 и 1, а число столбцов $m = \sum_{i=1}^{p} l_i$, т. е. равно суммарному количеству категорий для всех признаков $x^{(i)}$ ($i = \overline{1, p}$).

Таким образом, в отличие от матрицы $\mathbf{X}$ объекту соответствует строка матрицы $\mathbf{Y}$, а категориям переменных — столбцы. (Это не имеет принципиального значения, но упрощает обозначения.)

Матрица $\mathbf{Y}$ может быть представлена как объединение матриц $\mathbf{Y}_i$ с n строками и l_i столбцами, соответствующих бинарным представлениям признаков $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_p]$. Сумма элементов матрицы $\mathbf{Y}$ равна $y \ldots n \times p$.

17.2.2. Подход, основанный на непосредственном использовании матрицы Y. Матрицу $\mathbf{Y}$ можно рассматривать как таблицу с неотрицательными элементами с $l_1 = n$ строками, $l_2 = m$ столбцами и применить к ней АС из § 17.1.

С этой целью сначала получим аналог матрицы

$$\mathbf{F} = \frac{1}{y}\mathbf{Y} \tag{17.14}$$

Сумма элементов матрицы $\mathbf{F}$ равна 1. Сумма элементов любой строки этой матрицы (т. е любого объекта в данном случае) будет одинакова $f_i = 1/n$ ($i = 1, \ldots, n$), поскольку для любого объекта реализуется одна и только одна категория каждой переменной Следовательно, строки матрицы $\mathbf{F}$ имеют одинаковый вес, а матрица $\mathbf{D}_1 = \frac{1}{n}\mathbf{I}_n$, где $\mathbf{I}_n$ — единичная матрица размерности n. Сумму элементов для столбца матрицы $\mathbf{F}$, отвечающего k-й категории h-го признака, обозначим через

$$f_{\cdot hk} = n_k^h / np, \tag{17.15}$$

где n_k^h — число объектов, у которых h-й признак принял k-ю категорию Здесь для обозначения столбца используем два индекса· h и k, чтобы было более ясно, о какой категории идет речь Величины f_{hk} являются диагональными элементами матрицы $\mathbf{D}_2$ Далее будем также использовать диагональную матрицу $\mathbf{D} = np\,\mathbf{D}_2$, т. е. ее диагональные элементы суть частоты n_k^h.

Теперь можно определить профили строк (объектов) и столбцов (категорий) и ввести χ^2-метрики в пространствах объектов и категорий (см. п. 17.1.1).

Расстояние между k-й категорией h-го признака и l-й категорией r-го признака будет задаваться выражением

$$a^2(h_k, r_l) = \begin{cases} 0, & \text{если } h = r, \quad k = l; \\ n/n_k^h + n/n_l^r, & \text{если } h = r, \quad k \neq l, \\ n/n_k^h - 2n \cdot n_{kl}^{hr}/n_l^h n_l^r + n/n_l^r, & \text{если } h \neq r; \end{cases} \quad (17.16)$$

где n_{kl}^{hr} — число объектов, принявших категорию k для h-го и категорию l для r-го признаков

Расстояния между профилями строк (объектов) F_i и F_j в метрике χ^2 будут

$$d_{ij}^2 = \sum_{h=1}^{p} \sum_{k=1}^{l_h} (nf_{ih}^h - nf_{jk}^h)^2 / f_{hk} =$$

$$= \frac{1}{p} \sum_{h=1}^{p} \sum_{k=1}^{l_h} \omega_k^h (y_{ik}^h - y_{jk}^h)^2 = \frac{n}{p} (Y_i - Y_j)' \mathbf{D}^{-1} (Y_i - Y_j),$$

(17.17)

вес $\omega_k^h = n/n_k^h$, а величина n_l^h определена в (17.15) и является частотой k-й категории h-й переменной, y_{ik}^h (f_{ik}^h) — это соответствующий i-й строке и j-му столбцу элемент матрицы Y_h (F_h), Y_i — i-я строка матрицы $\mathbf{Y}$

Расстояние d_{ij} можно рассматривать как взвешенное (по категориям) хэммингово расстояние между объектами в пространстве бинарных переменных. Вес ω_k^h увеличивает вклад различий объектов по редким (по частоте) категориям

17.2.3. **Присвоение числовых меток объектам и категориям** (оцифровка). Действуя так же, как в п. 17.1.2, получим матрицу

$$\mathbf{\Phi} \quad \mathbf{D}_1^{-1/2} \mathbf{F} \mathbf{D}_2^{-1/2} = \frac{1}{\sqrt{p}} \mathbf{Y} \mathbf{D}^{-1/2} \quad (17.18)$$

и матрицы

$$\begin{cases} \mathbf{T}_r = \mathbf{\Phi}' \mathbf{\Phi} = \frac{1}{p} \mathbf{D}^{-1/2} \mathbf{Y}' \mathbf{Y} \mathbf{D}^{-1/2} \text{ (размера } m \times m); \\ \mathbf{T}_c = \mathbf{\Phi} \mathbf{\Phi}' = \frac{1}{p} \mathbf{Y} \mathbf{D}^{-1} \mathbf{Y}' \text{ (размера } n \times n). \end{cases} \quad (17.19)$$

Пусть теперь $\mu_0 \geqslant \mu_1 \geqslant \ldots \geqslant \mu_{l+} > 0$ — ненулевые собственные числа матрицы $\mathbf{T}_r$ ($\mathbf{T}_c$), а U_k (V_k) — соответст-

вующие им собственные векторы. Введем наборы числовых меток ($\mathbf{z}^{(k)}$, C_k) для строк (объектов) и столбцов (категорий):

$$\begin{cases} \mathbf{z}^{(k)} = \sqrt{\mu_k}\,\mathbf{D}_1^{-1/2} V_k = \sqrt{\mu_k\, n}\; V_k; \\ C_k = \sqrt{\mu_k}\,\mathbf{D}_2^{-1/2} U_k = \sqrt{\mu_k\, np}\;\mathbf{D}^{-1/2} U_k. \end{cases} \tag{17.20}$$

Вектор $\mathbf{z}^{(k)}$ будет n-компонентным вектором, а вектор C_k является m-компонентным. Так же как и в п. 17.1.3, k-й вектор (набор) меток для строк пропорционален вектору, компоненты которого равны проекциям нормированных профилей строк на k-й собственный вектор матрицы $\mathbf{T}_r$, т. е это вектор k-х главных компонент для профилей строк (в метрике χ^2) Аналогичное утверждение имеет место и для векторов меток для столбцов (категорий).

Таким образом, имеем l^+ наборов ($\mathbf{z}^{(k)}$, C_k) числовых меток для объектов и категории Иными словами, можно сказать, что использование МАС для обработки матрицы данных с переменными, измеренными в неколичественной шкале, приводит в результате к *квантификации* (или *оцифровке*) матрицы данных. Далее полученные наборы меток можно использовать для обработки данных как измеренных в количественных шкалах. Рассмотрим сначала, какими свойствами обладают наборы числовых меток, получаемые в МАС.

1. Существуют *тривиальные наборы* меток $\mathbf{z}^{(0)}$, C_0, соответствующие максимальному собственному числу $\lambda_0 = 1$. Все компоненты этих наборов равны 1. Причина появления наборов обсуждалась в п. 17.1.4.

2. Наборы меток для объектов $\mathbf{z}^{(1)}$, ..., $\mathbf{z}^{(l+)}$ можно рассматривать как новые количественные переменные (факторы). Эти переменные центрированы ($\sum_{i=1}^{n} z_i^{(j)} = 0$) с дисперсией $\frac{1}{n}\sum \left(z_i^{(j)}\right)^2 = \mu_j$ и попарно некоррелированы ($\sum_{i=1}^{n} z_i^{(j)} z_i^{(k)} = 0$, если $k \neq j$).

3. Используя уравнения (17.11), можно показать, что метки для объектов и категорий удовлетворяют следующим уравнениям перехода:

$$\begin{cases} \mathbf{z}^{(k)} = \dfrac{1}{p}\,\dfrac{1}{\sqrt{\mu_k}}\, Y C_k; \\ C_k = \dfrac{1}{\sqrt{\mu_k}}\,\mathbf{D}^{-1} Y'\,\mathbf{z}^{(k)}. \end{cases} \tag{17.21}$$

Следовательно, координата i-го объекта для k-го набора (т. е. значение k-го фактора для i-го объекта) *пропорцио-*

нальна среднему арифметическому значению меток категорий, реализующихся для этого объекта (всего реализуется p категорий, по одной для каждой переменной)

$$z_i^{(k)} = \frac{1}{p} \frac{1}{\sqrt{\mu_k}} \sum_{h=1}^{p} \sum_{j=1}^{l_h} y_{ij}^h C_{jk}^h, \qquad (17.22)$$

где C_{jk}^h ($j = 1, l_h$) — компоненты вектора C_k, соответствующие переменной $x^{(k)}$.

Аналогично координата (метка) j-й категории h-й переменной для k-го набора пропорциональна с множителем $(1/\sqrt{\mu_k})$ среднему значению фактора $z^{(k)}$ для объектов, имеющих значением для h-и переменной j-ю категорию, т. е. эта метка пропорциональна соответствующему *условному среднему значению* фактора $z^{(k)}$.

Из свойства 3 следует, что можно одновременно отображать объекты и категории в одной и той же системе координат для визуального анализа, так как метки, соответствующие объектам и категориям, измерены в одинаковых шкалах.

Пусть, например, следует провести визуальный анализ данных для первых двух факторов. Тогда для i-го объекта имеем координаты $z_i^{(1)}$, $z_i^{(2)}$, а для отображения категорий следует взять координаты $(\sqrt{\mu_1} c_{j1}^h, \sqrt{\mu_2} c_{j2}^h)$ (j-я категория h-й переменной).

17.2.4. Матрица Берта. Матрица $\mathbf{B} = \mathbf{Y}'\mathbf{Y}$, которая появляется в МАС, впервые была получена в работе [200] и носит название матрицы Берта (Burt). Это симметричная матрица, состоит из p^2 блоков. Имеется p диагональных блоков-матриц. Диагональный блок $\mathbf{Y}_j'\mathbf{Y}_j$ соответствует j-й переменной ($j = \overline{1, p}$) и представляет собой диагональную матрицу размера $l_j \times l_j$, так как две категории одной переменной не могут появляться одновременно. Диагональная матрица $\mathbf{D}$ имеет те же самые диагональные элементы, что и матрица $\mathbf{B}$.

Внедиагональный блок $\mathbf{Y}_j'\mathbf{Y}_i$ ($j \neq i$) представляет собой частотную таблицу сопряженности j-й и i-й переменных.

17.2.5. Подход, основанный на максимизации статистического критерия. Здесь рассматривается подход конструирования количественных факторов (переменных), которые наилучшим образом объясняют (аппроксимируют) матрицу $\mathbf{Y}$ в смысле некоторого статистического критерия. Одновременно с получением значений (метод для объектов) факторов при данном подходе получаются и метки для категорий пере-

менных. Получаемые метки, а также возникающие здесь метрики совпадают с метками и метриками, определяемыми на основе подхода, рассмотренного в п. 17.2.2, 17.2.3.

Пусть теперь следует присвоить числовые метки $\mathbf{v} = (v_1, \ldots, v_n)'$ n объектам. Потребуем, чтобы для набора меток $\mathbf{v}$ выполнялись условия центрирования и нормировки

$$\sum_{i=1}^{n} v_i = 0, \quad \sum_{i=1}^{n} v_i^2 = n. \tag{17.23}$$

Присвоить каждому объекту X_i некоторое числовое значение v_i — это и значит ввести некоторый новый признак (фактор) v.

Введем теперь величину (статистический критерий), определяющую качество набора меток

$$Q^2 = \sum_{j=1}^{p} R^2_{v,\,j}, \tag{17.24}$$

где $R^2_{v,\,j}$ — квадрат коэффициента множественной корреляции между фактором v и бинарными переменными y_k^j ($k = 1, \ldots, l_j$), т. е.

$$R^2_{v,\,j} = \frac{1}{n}\,\mathbf{v}'\,\mathbf{Y}_j\,(\mathbf{Y}_j'\,\mathbf{Y}_j)^{-1}\,\mathbf{Y}_j'\,\mathbf{v}. \tag{17.25}$$

Построенный из условия максимума Q^2 фактор v можно рассматривать как аналог первой главной компоненты, максимизирующей сумму квадратов коэффициентов корреляции (см. гл. 13).

Так как бинарные переменные линейно независимы, то

$$R^2_{v,\,j} = \sum_{k=1}^{l_j} r^2(v,\, y_k^j),$$

т. е. $R^2_{v,\,j}$ есть просто сумма квадратов корреляций бинарных переменных y_k^j, соответствующих переменной $x^{(j)}$, с v.

Будем теперь искать фактор из условия максимума критерия (17.24). Это приводит, с учетом условий нормировки (17.23), к следующей задаче на обобщенные собственные значения

$$\left(\mathbf{P} - \frac{p}{n}\,\mathbf{1}\mathbf{1}'\right)\mathbf{v} = \gamma\mathbf{v}, \tag{17.26}$$

где $\mathbf{1}$ — вектор размерности n с единичными компонентами; матрица $\mathbf{P} = \sum_{j=1}^{p} \mathbf{P}_j$; матрица $\mathbf{P}_j = \mathbf{Y}_j\,(\mathbf{Y}_j'\,\mathbf{Y}_j)^{-1}\,\mathbf{Y}_j^{-1}$

$= (p^j_{ik})$ $(i,\ k = \overline{1,\ n})$; элементы матрицы $\mathbf{P}_j$ вычисляются следующим образом:

$$p^j_{ik} = \begin{cases} 1/n^j_s, & \text{если } x^{(j)}_i = x^{(j)}_k = s; \\ 0, & \text{если } x^{(j)}_i \neq x^{(j)}_k. \end{cases} \tag{17.27}$$

где s — номер категории признака; i, k — номера объектов; n^j_s — число объектов, соответствующих s-й категории признака $x^{(j)}$. Суммарная матрица $\mathbf{P} = (p_{ik})$ есть матрица связей (близостей) между объектами X_i и X_j, измеряемых скалярными произведениями профилей строк в метрике χ^2. Каждая из величин $p_{ik} = Y_i \mathbf{D}^{-1} Y_j$ представляет собой сумму весов $1/n^j_s$ значений признаков $x^{(j)}$ $(j = \overline{1,\ p})$, которые совпадают для объектов i и k. Легко проверить, что матрица $\mathbf{P} = p\mathbf{T}_c$ (17.19). Ее собственные векторы совпадают с собственными векторами матрицы $\mathbf{T}_c$, и, следовательно, решением максимизационной задачи (17.24) будут факторы $z^{(1)}, \ldots, z^{(l)}$, определенные в п. 17.2.3.

17.2.6. Некоторые вопросы вычислительной реализации и интерпретации в множественном анализе соответствий

Итерационная вычислительная процедура. Факторы $(\mathbf{z}^{(k)}, C_k)$ можно получить, основываясь на решении проблемы собственных чисел и векторов для матриц $\mathbf{T}_c$, $\mathbf{T}_r$. Естественно, следует выбирать матрицу с минимальной размерностью, а сопряженные наборы меток для объектов и категорий получать с помощью уравнений перехода (17.21). Этот подход пригоден, когда какая-либо из матриц $\mathbf{T}_c$ или $\mathbf{T}_r$ помещается в оперативной памяти ЭВМ. В этом случае можно использовать и методы сингулярного разложения матриц, применяя их к матрице $\mathbf{\Phi}$ (17.18). При задачах большей размерности можно использовать итерационную процедуру, основываясь непосредственно на уравнениях (17.21). Так, используя уравнения (17.20), получим следующую процедуру:

$$\begin{cases} V^{(t+1)}_k = \mathbf{Y}\mathbf{D}^{-1/2} U^{(t)}_k; \\ U^{(t+1)}_k = \mathbf{D}^{-1/2}\, \mathbf{Y}'\, V^{(t+1)}_k; \\ \lambda^{(t)}_k = \| U^{(t+1)}_k \|, \ \mu^{(t)}_k = \lambda^{(t)}_k / p; \\ U^{(t+1)}_k = U^{(t+1)}_k / \lambda^{(t)}_k, \end{cases}$$

где t — номер итерации; k — номер фактора; λ_k — текущая оценка собственного числа.

Векторы $V_k^{(t)}$ на каждой итерации необходимо нормировать и центрировать в соответствии с условиями (17.23), а вектор $U_k^{(t)}$ нормировать.

Через определенное число итераций необходимо ортогонализировать текущие векторы $V_k^{(t)}$ (или $U_k^{(t)}$) к ранее найденным векторам $V_1, \ldots, V_{k-1}$ ($U_1, \ldots, U_{k-1}$) и тривиальным факторам.

Основной прием, делающий эту процедуру достаточно эффективной даже при больших размерностях m, n, связан с использованием специфической бинарной формы матрицы $\mathbf{Y}$. Действительно, умножение i-й строки матрицы на вектор $C_k^{(t)} = \mathbf{D}^{-1/2} U_k^{(t)}$ на самом деле требует использования только p *операций сложения* (поскольку только p элементов этой строки равно 1, а остальные равны 0). Таким образом, умножение матрицы $\mathbf{Y}$ на вектор $C_k^{(t)}$ требует всего $n \times p$ сложений (так же как и умножение $\mathbf{Y}'$ на $V_k^{(t)}$). Операция сложения намного экономнее по времени выполнения, чем операция умножения, и этих операций нужно всего $n \times p$ на каждой итерации, что и обеспечивает приемлемую эффективность вычислительной процедуры даже при больших размерностях n и p. При этом матрица $\mathbf{Y}$ может считываться поблочно из внешней памяти.

Итерационная процедура тем более пригодна, что обычно требуется небольшое количество факторов $q \ll m$. Существуют способы повышения эффективности итерационной процедуры, например одновременная итерация сразу нескольких векторов, и др. (см. [263]).

Собственные числа $\lambda_1, \ldots, \lambda_q$, полученные в результате итеративного процесса (17.28), будут связаны с собственными числами μ матриц $\mathbf{T}_r$, $\mathbf{T}_c$ соотношением $\lambda_k = p\mu_k$, а векторы U_k, V_k — совпадать с собственными векторами этих матриц. Используя соотношения (17.20), отсюда нетрудно перейти и к факторам $\mathbf{z}^{(k)}$, C_k.

Некоторые вопросы интерпретации. Как и при анализе главных компонент, перед исследователем, использующим МАС, возникает ряд вопросов, среди которых основными являются следующие: сколько факторов использовать и как их интерпретировать. Решение первого из них наталкивается на трудности, которых нет в анализе главных компонент, где наиболее принятый способ отбора числа значимых факторов связан с использованием доли следа ковариационной (корреляционной) матрицы, объясненной первыми q факторами (см. гл. 13). В МАС этот подход использовать обычно нельзя. Действительно, след матрицы $\mathrm{Sp}(\mathbf{T}_r) = \mathrm{Sp}(\mathbf{T}_c) =$

$= m/p$. Исключая вклад собственного числа $\mu_0 = 1$, соответствующего тривиальному фактору, имеем, что сумма ненулевых собственных чисел

$$\sum_{k=1}^{l^+} \mu_k = m/p - 1.$$

С другой стороны, величина $\mu_k \leqslant 1$ ($k = \overline{1, l^+}$). Поэтому доля следа, объясненная первыми q факторами, равная

$$\frac{\sum_{k=1}^{q} \mu_k}{(m/p - 1)} \leqslant qp/(m-p), \qquad (17.28)$$

может быть очень невелика, если общее число градаций m значнтельно.

Одна из возможностей эвристической оценки числа факторов состоит в сравнении собственных чисел с их средним значением — отбираются только факторы с собственными числами, большими среднего значения. Чтобы получить оценку среднего значения, оценим число положительных собственных чисел l^+. Она получается на основе следующих соображений, поскольку для каждой из матриц $\mathbf{Y}_k$, составляющих матрицу $\mathbf{Y}$, сумма ее столбцов равна вектору с единичными компонентами, ранг матрицы $\mathbf{Y}$ не более чем

$$\text{rank}\,(\mathbf{Y}) \leqslant m - p + 1. \qquad (17.29)$$

(Более точно $\text{rank}\,(\mathbf{Y}) < \min\,(n, m - p + 1)$, но поскольку обычно $n > m - p + 1$, можно использовать (17.29).) Поэтому для числа l^+ положительных собственных чисел для нетривиальных факторов верно неравенство $l^+ \leqslant m - p$. Отсюда в качестве оценки средней величины ненулевого собственного числа получаем

$$\bar{\mu} \geqslant \frac{\text{Sp}\,(\mathbf{T}_r) - 1}{m - p} = 1/p.$$

Интерпретация факторов. Подход к интерпретации выделенных факторов $v^{(1)}, \ldots, v^{(k)}$ ($z^{(1)}, \ldots, z^{(k)}$) основан на анализе множественных коэффициентов корреляции $R^2_{v^{(j)}, t}$ между факторами $v^{(j)}$ ($j = \overline{1, q}$) и исходными переменными (наборами бинарных переменных y_r^i). Аналогично интерпретации факторов в анализе главных компонент эти величины играют роль нагрузок переменных на факторы (см гл. 13). Из тех же соображений полезными для интерпретации являются коэффициенты корреляции между фактором

и бинарными переменными (категориями). Указанные величины, полезные для интерпретации факторов, получаются следующим образом:

квадрат коэффициента множественной корреляции между фактором $v^{(k)}$ и бинарными переменными $y_1^i, \ldots, y_{l_i}^i$

$$R^2_{v^{(k)}, i} = \frac{\lambda_k}{n} \sum_{j=1}^{l_i} u_{jk}^2;$$

квадрат коэффициента корреляции между фактором $v^{(k)}$ и бинарной переменной

$$r^2\left(v^{(k)}, y_j^i\right) = \frac{\lambda_k}{n} u_{jk}^2.$$

17.3. Алгоритмы оцифровки неколичественных переменных

Общие принципы. Пусть имеется матрица данных **X** из n p-мерных объектов, у которых все или часть признаков измерены в какой-либо неколичественной шкале — шкале порядка, номинальной и т.д.

Рассмотрим подход, позволяющий распространить на данные такого вида методы многомерного статистического анализа: анализа главных компонент, регрессионного, дискриминантного, кластер-анализа и т. д. Суть подхода заключается в оцифровке неколичественных переменных, т. е. в присвоении категориям неколичественных переменных «разумных», в рамках решаемой задачи, числовых меток. Этот же подход пригоден и для преобразования количественных переменных, которые предварительно подвергаются квантованию, и для анализа переменных смешанной природы. Метод приписывания меток для случая только неколичественных переменных приведен в [10, гл. 12]. Здесь формулируются критерии, подходящие для оцифровки с дальнейшим использованием преобразованной матрицы в различных видах анализа, а метод из [10, гл. 12] обобщается на случай данных смешанной природы.

Критерии, на основе которых производится присвоение числовых меток, зависят от используемого метода статистического анализа. Однако все они представляют собой некоторые функционалы матрицы ковариаций (корреляций) в пространстве оцифрованных признаков. Это связано прежде всего с тем, что матрица ковариаций (корреляций) является основным объектом, который используется методами статистического анализа.

Введем теперь некоторые очевидные требования, которым должны удовлетворять наборы числовых меток, получаемые в результате работы процедуры оцифровки. Пусть x — некоторый неколичественный признак из матрицы данных $\mathbf{X}$, имеющий l_x градаций (категорий) значений. Пусть каждой из l_x градаций присвоена числовая метка $c_j^{(x)}$ ($j = \overline{1, l_x}$). Поскольку корреляции между признаком x и другими признаками не зависят от преобразования сдвига и масштабирования меток, потребуем выполнения условий центрированности и нормировки

$$\sum_{i=1}^{n} c_{r(i)}^{(x)} = 0; \quad \frac{1}{n} \sum_{i=1}^{n} c_{r(i)}^{(x)\,2} = 1, \tag{17.30}$$

где $r(i)$ — номер градации признака x для i-го объекта.

Пусть теперь $\hat{f}_{\cdot i} = n_i^x/n$ — частота i-й градации признака x у объектов из $\mathbf{X}$. Тогда условия (17.30) можно эквивалентным образом записать в виде

$$\frac{1}{n} \sum_{r=1}^{l_x} n_i^x c_r^x = 0; \quad \frac{1}{n} \sum_{r=1}^{l_x} n_i^x c_r^{(x)\,2} = 1. \tag{17.30'}$$

Выполнение условий (17.30), (17.30′) гарантирует, в частности, от появления тривиальных наборов меток, когда числовые метки, присваиваемые градациям признака x, одинаковы.

Оцифровка для сокращения размерностей, статистического исследования зависимостей, кластер-анализа. В этом случае категориям неколичественных признаков приписываются числовые метки, удовлетворяющие условиям (17.30) и максимизирующие величину

$$Q^2 = \sum_{i<j}^{p} \rho_{ij}^2, \tag{17.31}$$

где $i, j = \overline{1, p}$; p — число признаков; ρ_{ij} — коэффициенты корреляции между i-м и j-м признаками после кодировки.

Пусть теперь множество переменных $x^{(1)}, \ldots, x^{(p)}$ разбито на две группы — группу $X^{(1)}$ из q переменных, подлежащих кодировке (оцифровке), и группу $X^{(2)}$ из $p - q$ переменных, для которых сохраняется исходная шкала (или исходные значения меток). В частности, в группе $X^{(2)}$ могут быть переменные, измеренные и в количественной шкале. Для определенности будем считать, что признаки пронумерованы так, что в $X^{(1)}$ входят признаки $x^{(1)}, \ldots, x^{(q)}$, а в $X^{(2)}$ — при-

знаки $x^{(q+1)}$, ..., $x^{(p)}$. Критерий Q^2 может быть представлен в виде суммы трех слагаемых: $Q^2 = Q_1^2 + Q_{1,2}^2 + Q_2^2$, где Q_1^2 — сумма квадратов коэффициентов корреляции переменных из $X^{(1)}$; $Q_{1,2}^2$ — сумма квадратов коэффициентов корреляции между переменными из $X^{(1)}$ и $X^{(2)}$; Q_2^2 — сумма квадратов коэффициентов корреляции между переменными из $X^{(2)}$. Величина слагаемого Q_2^2 не зависит от кодировки, поэтому определение оптимальных меток будем проводить исходя из условия максимума критерия $\widetilde{Q}^2 = Q_1^2 + Q_{1,2}^2$.

Приведем теперь формулы для вычисления оценок коэффициентов корреляции, входящих в состав сумм Q_1^2 и $Q_{1,2}^2$. Пусть признаки $x^{(i)} \in X^{(1)}$ и $x^{(j)} \in X^{(1)}$ и пусть l_i — число категорий признака $x^{(i)}$. Тогда, если выполнены условия нормировки (17.30′), получаем, что

$$\widehat{\rho}_{ij} = C_i' \widehat{\mathbf{F}}(i, j) C_j,$$

где C_i — вектор числовых меток для категорий признака $x^{(i)}$, $\widehat{\mathbf{F}}(i, j)$ — нормированная таблица сопряженности размера $l_i \times l_j$ между признаками $x^{(i)}$ и $x^{(j)}$, т. е. $\widehat{\mathbf{F}}(i, j) = \frac{1}{n} \mathbf{N}(i, j)$ (см. п. 17.1.1).

Пусть теперь признак $x^{(i)} \in X^{(1)}$, а $x^{(j)} \in X^{(2)}$ и пусть предварительно признак $x^{(j)}$ нормирован и центрирован. Тогда $\widehat{\rho}_{ij} = C_i' \widehat{\mathbf{P}}_i \overline{C}_j^{(i)}$, где $\widehat{\mathbf{P}}_i = \operatorname{diag}(\widehat{p}_1, \ldots, \widehat{p}_{l_i})$, $\widehat{p}_k$ — частота появления k-й градации признака $x^{(i)}$ $(k = \overline{1, l_i})$; $(\overline{C}_j^{(i)})' = (\overline{c}_{1/}^{(i)}, \ldots, \overline{c}_{l_i/}^{(i)})$, а $\overline{c}_k^{(i)}$ — среднее значение признака $x^{(j)}$ на множестве объектов с k-й категорией признака $x^{(i)}$ $(k = \overline{1, l_i})$. Для каждого признака $x^{(i)} \in X^{(1)}$ введем симметричную неотрицательно определенную матрицу $\mathbf{A}_i$, такую, чтобы удовлетворялось равенство $\partial \widetilde{Q}/\partial C_i = \mathbf{A}_i C_i$.

Непосредственным дифференцированием получаем, что

$$\mathbf{A}_i = \sum_{\substack{j \neq i \\ (x \in X^{(1)})}}^{q} \mathbf{N}(i, j) C_j C_j' \mathbf{N}(j, i) + \sum_{\substack{j=q+1 \\ (x \in X^{(2)})}}^{p} \widehat{\mathbf{P}}_i \overline{C}_j^{(i)} (\overline{C}_j^{(i)})' \widehat{\mathbf{P}}_i.$$

Вычислительная процедура. Числовые метки, максимизирующие величину критерия $\widetilde{Q}_2$, находятся в результате

итерационного процесса, аналогичного описанному в [11, гл. 12].

Пример 17.4. [66] Рассмотрим применение метода оцифровки по критерию (17.31) к данным табл. 17.2, представляющей результаты наблюдения за 12 посетителями кафе (пример условный). Переменные имеют следующий смысл:

$x^{(1)}$ — сумма, затраченная посетителем, ден. ед.;

$x^{(2)}$ — время, проведенное посетителем в кафе, мин.;

$x^{(3)}$, $x^{(4)}$, $x^{(5)}$ — соответственно закуска, блюдо и напиток, выбранные посетителем.

Таблица 17.2

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
100	63	1	4	1	125	135	2	3	4
85	63	1	2	1	170	95	2	1	3
65	45	1	2	2	180	135	2	1	4
65	45	2	2	2	95	63	3	4	1
110	95	2	3	3	105	95	3	3	3
120	95	2	3	3	175	135	3	4	4

Переменные $x^{(1)}$ и $x^{(2)}$ — количественные, а $x^{(3)}$, $x^{(4)}$, $x^{(5)}$ — номинальные категоризованные, переменная $x^{(3)}$ имеет три, а $x^{(4)}$ и $x^{(5)}$ — по четыре градации.

Возможно использование переменных, которые не будут подвергаться оцифровке, но их вклад в критерий (17.31) будет учитываться. В данном примере это количественные переменные $x^{(1)}$, $x^{(2)}$. Ниже приводятся результаты применения оцифровки процедуры.

МАТРИЦА КОРРЕЛЯЦИЙ ДО ПРЕОБРАЗОВАНИЯ

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
$x^{(1)}$	1.0000	0.8514	0.3846	—0.1694	0.7244
$x^{(2)}$	0.8514	1.0000	0.4474	0.0067	0.8847
$x^{(3)}$	0.3846	0.4474	1.0000	0.3441	0.4228
$x^{(4)}$	—0.1694	0.0067	0.3441	1.0000	—0.1940
$x^{(5)}$	0.7244	0.8847	0.4228	—0.1940	1.0000

СУММА КВАДРАТОВ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ 2.0191

МАТРИЦА КОРРЕЛЯЦИЙ ПОСЛЕ ПРЕОБРАЗОВАНИЯ

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
$x^{(1)}$	1.0000	0.8514	—0.4969	—0.7737	—0.8451
$x^{(2)}$	0.8514	1.0000	—0.5685	—0.7319	—0.9719
$x^{(3)}$	—0.4969	—0.5685	1.0000	0.6288	0.6276
$x^{(4)}$	—0.7737	—0.7319	0.6288	1.0000	0.8200
$x^{(5)}$	—0.8451	—0.9719	0.6276	0.8200	—1.0000

СУММА КВАДРАТОВ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ 4.8241

ТАБЛИЦА НАЙДЕННЫХ МЕТОК

	$x^{(3)}$	$x^{(1)}$	$x^{(5)}$
1	0.8880	—0.5079	0.6163
2	—0.4589	0.6095	0.5304
3	—0.0274	—0.4869	0.5086
4	0.0000	0.3784	—0.2891

Из сравнения матриц корреляций до и после оцифровки следует, что после оцифровки значения некоторых коэффициентов корреляции значительно возросли по абсолютной величине. Так, величина ρ_{42} до оцифровки была 0.0067, а после оцифровки стала равной —0.7319.

Оцифровка для линейного дискриминантного анализа. Для задач классификации оцифровка неколичественных признаков производится по критерию, предложенному в [68]. Этот критерий построен на том, что основной информацией, которую используют линейные дискриминантные функции для классификации, являются различия средних значений признаков в разных классах, измеренные в единицах дисперсии (см. гл. 1). Другие компоненты информации о различиях между распределениями классов используются линейной дискриминантной функцией в меньшей степени. Исходя их этого в качестве набора числовых меток для категорий некоторого признака x примем числа, максимизирующие сумму оценок квадратов расстояний Махаланобиса от общего центра тяжести по признаку x ($m(C)$) до центров классов по этому же признаку ($m_j(C)$, $j = 1,\ldots, k$):

$$t^2(C) = \sum_{j=1}^{k} \alpha_j (m_j(C) - m(C))^2 / \sigma^2(C), \qquad (17.32)$$

где α_j — вероятность появления объектов из j-го класса; k — число классов; $\sigma^2(C)$ — усредненная дисперсия.

Введем в рассмотрение таблицу сопряженности **F**, столбцы которой соответствуют категориям классификационной

переменной, а строки — категориям признака x. Элемент f_{ij} ($i=1, \ldots, l_x$; $j=1, \ldots, k$) является, таким образом, вероятностью появления i-й категории переменной x в j-м классе. (В реальной ситуации мы обычно имеем дело с обучающими выборками, и поэтому вместо частот известны лишь их оценки $\widehat{f_{ij}}$ — частоты категории i в классе j.)

Теперь величины $m(C)$, $m_i(C)$, $\sigma^2(C)$, входящие в (17.32), можно представить в следующем виде:

$$\alpha_j = \sum_{i=1}^{l_x} f_{ij} = f_{\cdot j};$$

$$m_j(C) = \sum_{i=1}^{l_x} c_i f_{ij}/f_{\cdot j} = (\mathbf{q}'C);$$

$$m(C) = \sum_{j=1}^{k} f_{\cdot j} m_j = \sum_{i=1}^{l_x} f_{i\cdot} c_i;$$

$$\sigma_j^2 = \sum_{i=1}^{l_x} (c_i - m_i(C))^2 f_{ij}/f_{\cdot j} = \sum_{i=1}^{l_x} c_i^2 f_{ij}/f_{\cdot j} - m_j^2(C);$$

$$\sigma^2(C) = \sum_{j=1}^{k} f_{\cdot j}\sigma_j^2 = \sum_{i=1}^{l_x} c_i^2 f_{i\cdot} - \sum_{j=1}^{k} f_{\cdot j} m_j^2(C).$$

Вводя матрицы

$$\mathbf{D}_1 = \operatorname{diag}(f_{1\cdot}, \ldots, f_{l_x\cdot}).$$

$$\mathbf{D}_2 = \operatorname{diag}(f_{\cdot 1}, \ldots, f_{\cdot k}),$$

мы можем записать

$$\rho^2 = \sum_{j=1}^{k} \alpha_j m_j^2 = C'\,\mathbf{F}\mathbf{D}_2^{-1}\,\mathbf{F}'\,C;\quad m(C) = \mathbf{D}_1 C;$$

$$s^2 = \sum_{i=1}^{l_x} f_{i\cdot} c_i^2 = C'\,\mathbf{D}_1 C.$$

В новых обозначениях критерий (17.32) можно записать в виде

$$t^2(C) = (\rho^2 - m^2(C))/(s^2 - \rho^2), \qquad (17.33)$$

Очевидно, задача поиска максимума $t^2(C)$ инвариантна относительно преобразований сдвига и масштаба координат

C, а потому может быть сведена к задаче на условный экстремум

$$\rho^2(C) \Rightarrow \max_C \tag{17.34}$$

при условиях $m(C) = 0$, $s^2 = 1$,
что приводит в результате к обобщенной задаче на собственные числа

$$(\mathbf{F}\mathbf{D}_2^{-1}\mathbf{F}' - \lambda \mathbf{D}_1) C = 0. \tag{17.35}$$

Но эта задача эквивалентна рассмотренной в п. 17.1.2 задаче на собственные числа $\mathbf{T}_c V = \lambda V$ с переходом $C = \mathbf{D}_1^{-1/2} V$. При этом, чтобы удовлетворить условиям (17.34), мы должны взять собственный вектор, соответствующий второму по величине собственному числу λ_1 (см. п. 17.1.2). Итак, мы снова пришли к каноническим меткам. Величина собственного числа $\lambda_1 = \rho^2(C)$ связана с отношением $t^2(C)$ выражением

$$t^2(C) = \rho^2(C)/(1 - \rho^2(C)). \tag{17.36}$$

Как показано в гл. 2, при объемах выборки, сравнимых с числом переменных p и числом градаций $l_{(x)}$, применение процедуры оцифровки следует проводить с осторожностью. В частности, целесообразно оцифровывать те признаки, для которых значение $\rho^2(C)$ статистически высоко значимо. Приведем один из полезных критериев, для определения допустимости оцифровки, основанный на асимптотическом распределении выборочных собственных чисел $\widehat{\lambda}_i$. Оказывается (см. [263, 265, 287]), имеет место следующий результат.

Пусть таблица сопряженностей $\mathbf{F}$ с l_1 строками и l_2 столбцами ($l_1 > l_2$) удовлетворяет условию независимости, т. е. собственные числа канонических уравнений $\lambda_1 = \ldots = \lambda_{l_2-1} = 0$ (существует только тривиальный набор меток для $\lambda_0 = 1$). Тогда выборочные числа $n\widehat{\lambda}_1, \ldots, n\widehat{\lambda}_{l_2-1}$ для $\widehat{\mathbf{F}}$ асимптотически распределены как собственные числа матрицы, подчиненной распределению Уишарта [16] $W(l_1 - 1, l_2 - 1, \mathbf{I}_{l_2-1})$, где $\mathbf{I}_{l_2-1}$ — единичная матрица размерности $l_2 - 1$.

Теперь для построения критерия можно воспользоваться, например, результатами по распределению максимального собственного числа матрицы Уишарта [241] в асимптотике Колмогорова (см. гл. 2).

Используя эти результаты, получаем, что при $l_1, l_2 \to \infty$, значение $n\widehat{\lambda}_1/(l_1 - 1)$ почти наверное сходится к величине $\tilde{\lambda}_1 = (1 + \sqrt{y})^2$, где $y = (l_2 - 1)/(l_1 - 1)$.

Это приводит к следующей формулировке критерия — переменную x следует использовать для линейной классификации в оцифрованном виде, если после оцифровки величина $\rho^2(C)$ будет удовлетворять неравенству

$$\rho^2(C) \geqslant \frac{l_x}{n}\left(1+\sqrt{y}\right)^2, \qquad (17.37)$$

где $y = (k-1)/(l_x-1)$.
В случае, когда $k > l_x$, следует поменять местами k и l_x в (17.37).

ВЫВОДЫ

1. АС (см. § 17.1) является аналогом метода главных компонент в пространстве профилей строк и столбцов ТС. В результате применения АС получаются сопряженные наборы числовых меток для строк и столбцов АС (оцифровки), что позволяет, в частности, получать визуальное отображение строк и столбцов ТС на диаграммах рассеивания. АС применим не только к ТС, но и к матрицам данных типа «объект — свойство», элементы которых неотрицательны и имеют одинаковую природу (например, доли, проценты, денежные платежи и т. д.).
2. МАС (см. § 17.2) применим к матрицам данных типа «объект — свойство» с переменными, измеренными в неколичественных категоризованных шкалах. Его применение позволяет ввести χ^2-метрику между объектами и между категориями переменных, а также получить наборы числовых меток для объектов (новые факторы) и для категорий. Введенная χ^2-метрика и факторы могут быть использованы дальше для реализации других процедур статистического анализа, в частности для визуального отображения объектов и категорий.
3. Оцифровку неколичественных переменных (метки для категорий) можно получить и на основе максимизации некоторых статистических критериев (см. § 17.3), выбираемых в зависимости от целей дальнейшего анализа (главные компоненты, дискриминантный анализ). Получаемые наборы меток часто близки к получаемым в МАС, но, вообще говоря, не совпадают с ними.

Раздел IV. РАЗВЕДОЧНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ И НАГЛЯДНОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ

В настоящем разделе рассматривается совокупность моделей и методов, позволяющих анализировать многомерные данные с помощью их отображения в пространство низкой размерности с сохранением существенных для исследователя их структурных особенностей.

В некоторых случаях структура данных оказывается такой сложной, что небольшого числа таких проекций будет недостаточно для их представления и возникает потребность описания этой структуры на основе агрегирования информации, содержащейся в достаточно большом числе таких низко размерных проекций. Типичной задачей такого класса является задача восстановления плотности многомерной случайной величины.

Такая ситуация имеет место при некоторых конфигурациях расположения классов в задаче дискриминантного анализа, когда число классов превышает размерность выборки, и т. п.

Рассматриваемые модели и методы, естественно, делятся на два класса, в зависимости от размерности q пространства, куда отображаются данные. Если $q = 1, 2$ (в крайнем случае 3), то они в первую очередь относятся к собственно разведочному анализу (РА), когда по некоторому критерию при помощи вычислительной процедуры оптимизации ищут отображения, дающие наиболее выразительные проекции, а окончательное решение принимается визуально путем анализа, например на экране дисплея, гистограмм отображенных данных $q = 1$ или их диаграмм рассеивания для $q = 2$. Здесь наибольший успех можно ожидать в задачах разделения смесей, кластеризации, т. е. когда ищется явно выраженная структура. Успеха можно достичь и в задачах обнаружения неинформативных признаков.

К этому же классу относятся модели и методы решения тех задач, когда исходным статистическим материалом является не описание объектов в виде вектор-признаков, а описа-

ние в виде наборов линейных функций от этих вектор-признаков. Типичной задачей этого вида являются задачи статистического анализа по результатам косвенных измерений. Если $q \ll p$, то соответствующие методы можно рассматривать как один из подходов к снижению размерностей, например для целей кластер-анализа.

В этом разделе в основном рассматриваются методы, связанные с линейным проецированием данных. Совокупность таких методов в последнее время получила большое развитие и известна в статистической литературе как «projection pursuit» (PP). Ряд других методов, которые можно отнести к РАД, рассмотрен в предыдущих главах. Это метод главных компонент, кластер-анализ, методы многомерного шкалирования, а в случае неколичественных переменных — анализ соответствий.

Глава 18. РАЗВЕДОЧНЫЙ АНАЛИЗ. ЦЕЛИ, МОДЕЛИ СТРУКТУР ДАННЫХ, МЕТОДЫ И ПРИЕМЫ АНАЛИЗА

18.1. Цели разведочного анализа и модели описания структуры многомерных данных

Разведочный анализ данных (РАД; Exploratory data analysis) употребляется, когда, с одной стороны, у исследователя имеется таблица многомерных данных, а с другой стороны, априорная информация о физическом (причинном) механизме генерации этих данных отсутствует или неполна. В этой ситуации РАД может оказать помощь в *компактном* и *понятном* исследователю описании структуры данных (например, в форме визуального представления этой структуры), отталкиваясь от которого он уже может «прицельно» поставить вопрос о более детальном исследовании данных с помощью того или иного раздела статистического анализа, обоснования полученной структуры данных с помощью аппарата проверки статистических гипотез, а также, возможно, сделать некоторые заключения и о причинной модели данных. Этот этап называется «подтверждающим анализом данных» (confirmatory data analysis). Иногда выявление структуры данных с помощью РАД может оказаться и завершающим этапом анализа. С другой стороны, ряд методов РАД можно рассматривать и как методы подготовки данных для последующей статистической обработки без ка-

кого-либо изучения структуры данных, которое предполагается осуществить на последующих этапах. В этом случае этап РАД играет роль некоторого этапа перекодировки и преобразования данных (путем, например, сокращения размерности) в удобную для последующего анализа форму. В любом случае, с какой бы целью ни применялись методы РАД, основная задача — *переход к компактному описанию данных при возможно более полном сохранении существенных аспектов информации*, содержащихся в исходных данных. Важно также, чтобы описание было понятным для пользователя. Впервые термин «разведочный анализ данных» был введен Дж. Тьюки в 1962 г.

Модели структуры многомерных данных. Пусть данные заданы в виде матрицы данных. Объекты можно представить в виде точек в многомерном (p-мерном) пространстве. Для описания структуры этого множества точек в РАД используется одна из следующих *статистических моделей*:

а) модель облака точек примерно эллипсоидальной конфигурации;

б) кластерная модель, т. е. совокупность нескольких «облаков» точек, достаточно далеко отстоящих друг от друга;

в) модель «засорения» (компактное облако точек и при этом присутствуют далекие выбросы);

г) модель носителя точек как многообразия (линейного или нелинейного) более низкой размерности, чем исходное; типичным примером является выборка из вырожденного распределения;

д) дискриминантная модель, когда точки разделены некоторым образом на несколько групп и дана информация о их принадлежности к той или иной группе.

В рамках модели (г) можно рассматривать и регрессионную модель, когда соответствующее многообразие допускает функциональное представление $X_{II} = F(X_I) + \varepsilon$, где X_I и X_{II} — две группы переменных из исходного набора (переменные из X_{II} носят тогда название прогнозируемых переменных, а из X_I — предсказывающих переменных); ε — ошибка предсказания.

Разумеется, реальные данные обычно лишь приближенно могут следовать этим моделям, более того, структура данных может не подходить ни под одну из указанных в описании моделей даже приближенно.

Модели описания структуры зависимостей. В пространстве переменных для описания структуры *зависимостей* между переменными часто используются следующие модели: модель независимых переменных, модель линейно зависи-

мых переменных, древообразная модель зависимости, факторная модель для линейно зависимых переменных, кластерная модель (произвольные коэффициенты связи), иерархическая модель зависимости.

Основные методические приемы при проведении разведочного анализа данных. Способы анализа и интерпретации результатов в значительной степени зависят от выбранного метода обработки. Однако можно выделить ряд эффективных приемов и подходов к анализу результатов, которые являются наиболее общими и в значительной степени определяют специфику собственно разведочного анализа, отличают его от остальных этапов статистической обработки. Это визуализация данных и манипуляции с данными на основе графического отображения; использование аппарата активных и иллюстративных переменных; преобразование данных, облегчающее выявление структур, анализ остатков.

18.2. Визуализация данных

18.2.1. Роль визуализации в разведочном анализе данных. Как выше указывалось, основное назначение РАД — дать компактное и понятное для исследователя описание структуры данных или структуры зависимости переменных. Визуализация данных, которая предполагает получение тем или иным способом их графического отображения, так что исследователь может просто путем непосредственного визуального анализа этого изображения определить, имеет ли место одна из моделей структуры данных (а, б, в, г), является, по-видимому, наиболее наглядным способом описания.

Графическое отображение (гистограммы, диаграммы рассеивания) может быть получено непосредственно в пространстве исходных переменных. Однако «информативное» графическое отображение многомерных данных получается с помощью методов РАД, нацеленных на выявление перечисленных структур данных и зависимостей (например, главных компонент, анализа соответствий, целенаправленного проецирования и т. д.). В результате применения этих методов получаются образы объектов, переменных и (для неколичественных переменных методом соответствий анализа) категория в виде точек обычно размерности 1—3. Выходная размерность данных может быть и больше 3, но для графического отображения все равно берутся какие-либо одна, две или три их координаты, обычно при этом первые координаты более информативны и используются для визуального анализа в первую очередь. Быстро возрастающая роль визуального

анализа многомерных данных стимулирована широким распространением и доступностью технических (вычислительных) средств, обеспечивающих построение визуальных образов. В 60-е и 70-е годы основным и наиболее широко использовавшимся техническим средством для представления графических форм, возникающих в статистическом анализе, служило алфавитно-цифровое печатающее устройство (АЦПУ). Существенно менее доступными были графопостроители и графические дисплеи. Тем не менее некоторые динамические формы визуального анализа были разработаны уже в начале 70-х годов именно с целью использования возможностей графического дисплея, обслуживаемого достаточно мощной ЭВМ. В качестве такого примера можно привести систему PRIME [230].

Современная графика для статистического анализа обладает всеми свойствами и преимуществами компьютерной графики — построение, обработка и модификация графических форм возможна в интерактивном режиме и за короткое время.

18.2.2. Диаграммы рассеивания. Рассмотрим вопросы визуализации многомерных данных, связанные с использованием диаграмм рассеивания (ДР), которые являются широко распространенной, простой и эффективной формой визуального представления данных. Некоторые другие формы визуального представления данных (гистограммы, графики оценок плотности и др.) рассмотрены в [223, 11, гл. 10]. В гл. 8 книги приведены формы визуализации структур, возникающих в иерархических процедурах кластер-анализа.

ДР многомерных данных является визуальной формой представления результатов некоторого отображения исходной матрицы данных в двумерное евклидово пространство. Роль исходной матрицы данных может играть матрица «объект — свойство» или матрица близостей (отношений «объект — объект», «переменная — переменная»). В качестве отображенных на ДР единиц могут выступать объекты, переменные, категории переменных (если переменные неколичественные). Далее они будут называться *отображенными единицами* (ОЕ). Графические же элементы, с помощью которых ОЕ изображаются на ДР, будут называться *выразительными элементами* (ВЭ). В табл. 18.1 приведены основные методы анализа, порождающие информативные ДР.

Рассмотрим теперь некоторые способы, позволяющие улучшить способность ДР к отображению структурных данных.

Маркирование ОЕ. Маркирование достигается, в зависимости от технических возможностей средств графического

Таблица 18.1

№ п/п	Типы матрицы данных	Цель визуального анализа	Способ получения отображения	Единицы, отображаемые на ДР
1	2	3	4	5
1	«Объект — признак» с колнчественными переменными	Выделения любой структуры в даиных; интерпретацня линейных и нелинейных факторов	Главные компоненты; нелинейное отображение (см. гл. 13)	Объекты, переменные
2	«Объект — признак» с колнчественнымн переменнымн илн с переменными смешанной природы	Выделение любой структуры в данных; интерпретация выделяемых структур	Множественный анализ соответствий, нелинейное отображение (см. гл 17, 13)	Объекты, категорни переменных, переменные, одновременное отображение объектов и переменных
3	«Объект — признак» с количествеиными переменными с добавленнем группирующей переменной	Изучение взаимного пространственного расположения групп объектов в дискриминантном или кластерном анализе	Канонические направления по Рао, целенаправленное проецирование при наличии обучающих выборок (см § 19 4)	Объекты, группы объектов (задаваемые различными маркерами для различных групп)
4	«Объект — признак»	Выделение аномальных наблюдений (outliers)	Целенаправленное проецированне (см. § 19.5)	Объекты с маркированием объектов, подозрительных как «outliers»
5	«Объект — признак»	Выделение кластеров	Кластер-анализ и далее отображения из стр. 3 табл.; целенаправленное проецирование (см. § 19 2, 19 3)	Объекты
6	«Объект — признак»	Выделение иелииейных структур	Целенаправленное проецирование (см § 19.6)	Объекты
7	«Объект — объект» или «признак — признак»	Выделение любой структуры в данных	Многомерное шкалирование; анализ соотвстствий (см. гл. 16, 17)	Объекты илн перемеиные, или категории в зависимости от типа матрицы

отображения, путем вариации окраски, формы и величины ВЭ, используемых для представления на ДР отображаемых единиц — объектов, переменных, категорий.

Так, обыденной практикой в дискриминантном и кластерном анализе является выделение на ДР, путем маркирования объектов, принадлежащих к разным группам, категорий, принадлежащих к разным переменным в множественном анализе соответствий.

Другой пример — маркирование объектов, подозрительных на аномальность, на ДР, используемой в целенаправленном проецировании для выделения аномальных наблюдений (см. пример 19.3).

Маркирование может быть использовано и с целью отображения на двумерной ДР информации о некотором дополнительном третьем измерении (например, о третьей главной компоненте на ДР, соответствующей двум первым ГК). Для этого, например, объекты изображаются точками, а из этих точек восстанавливается отрезок, параллельный оси $0y$ (вертикальной оси). Длина этого отрезка пропорциональна значению третьей координаты, а ее направление вверх или вниз соответствует знаку этой координаты. Если количество ОЕ невелико, то можно маркировать и четвертое измерение с помощью, например, горизонтальных отрезков. Другой возможностью на цветном графическом дисплее является использование окраски и ее интенсивности. Например, красная, оранжевая и желтая окраска для положительных значений третьей координаты (диапазон значений разбивается на три градации — большие, средние, малые) и синий, циан, белый — для отрицательных значений (с аналогичным разбиением диапазона отрицательных значений на три градации). Разумеется, такие ДР могут лишь частично передать информацию о взаимном расположении точек в пространстве более чем двух измерений, и Дж. Тьюки предлагает называть эти ДР $2\frac{1}{2}$-мерными [323].

Изменение масштаба. Меняя масштабы ДР по вертикали и горизонтали, тем самым изменяем метрику двумерного изображения — визуально наблюдаемые расстояния и взаимное расположение точек (изменение масштаба соответствует некоторому линейному преобразованию ОЕ в двумерном пространстве). Тем самым можно добиться более выраженного визуального представления тех или иных структур на ДР.

Один из простых технических приемов изменения масштабов состоит в следующем. Обычно при построении ДР задаются ее размеры — колнчество строк (линий) по оси

$0y$ и интервалов по оси $0x$. Размах значений ОЕ по оси $0y$ делится на число строк, а размах значений по оси $0x$ — на число интервалов. Полученные частные и являются масштабами измерений. Меняя задаваемое на ДР число строк и интервалов, можно добиться таким образом и изменения масштабов.

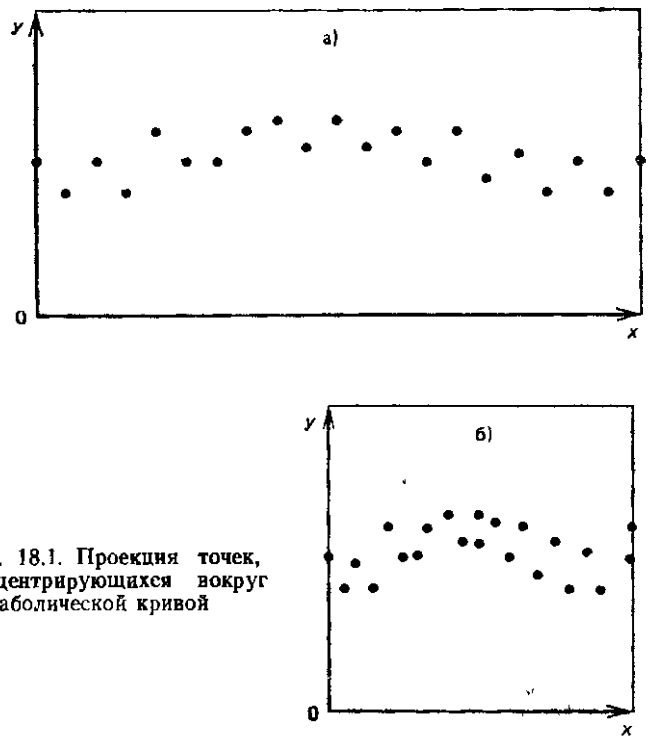


Рис. 18.1. Проекция точек, концентрирующихся вокруг параболической кривой

На рис. 18.1, а представлено облако точек, которые концентрируются вокруг некоторой кривой. Сжатие по оси $0x$ делает эту структуру более выраженной, что и демонстрирует рис. 18.1, б.

При построении ДР часто используются и нелинейные преобразования координат ОЕ, например логарифмический масштаб и т. д., что в ряде случаев позволяет выявить дополнительные структурные особенности в данных.

18.2.3. Динамические формы диаграмм рассеивания

Многооконные ДР. Новые возможности для визуального анализа представляет одновременное изучение нескольких ДР для одного и того же множества ОЕ. На экране дисплея

создается несколько окон, в каждом из которых высвечивается своя ДР. При этом отображения исходной матрицы данных могут быть получены как в рамках применения одного какого-либо статистического метода (например, главных компонент), так и при применении нескольких методов (например, целенаправленное проецирование для выделения кластерной структуры (см. § 19.4) и главных компонент (см. гл. 13)). Конечно, рассмотрение изображений на нескольких ДР полезно и в статическом режиме. Однако введение динамических элементов позволяет использовать качественно новые возможности [183, 315].

Простым, но эффективным приемом является использование подвижного окна, положение и размеры которого управляются пользователем. Окно движется по одной из ДР и ОЕ, попавшие внутрь этого окна, маркируются одновременно на всех ДР. Для каких целей может быть использовано подвижное окно? Приведем только некоторые возможные применения.

Одно из возможных использований — проверка предположения о том, что выделяемое сгущение ОЕ на какой-либо ДР действительно представляет собой кластер в исходном многомерном пространстве, а не является просто свойством данной проекции. Для этого подвижное окно накладывают на сгущение и наблюдают, как расположены те же самые точки на других ДР. Если на какой-либо ДР ВЭ, соответствующие выделенным с помощью подвижного окна ОЕ, разбросаны равномерно по всему экрану, то, значит, сгущение не является кластером. Если же на всех экранах выделенная совокупность ОЕ распределена компактно, уверенность в том, что полученное образование действительно некоторый кластер, возрастает. Конечно, ДР нужно выбирать так, чтобы расстояния между ОЕ на них были бы величины одного порядка.

Другое возможное использование состоит в изучении условных распределений. Действительно, фиксация точек внутри подвижного окна на какой-либо из ДР соответствует тому, что рассматриваем на других ДР распределение ОЕ, удовлетворяющих условиям $x_l < x < x_r$, $y_d < y < y_u$, где x_l, x_r, y_d, y_u — границы окна; x, y — координаты точек на ДР с подвижным окном. На рис. 18.2 показана ситуация, когда точки, достаточно равномерно распределенные внутри подвижного окна на левой ДР (рис. 18.2, а), концентрируются вокруг некоторой кривой линии на другой ДР (рис. 18.2, б).

Наконец, наиболее обыденный путь использования подвижного окна состоит в использовании его для идентифика-

ции ОЕ. Для этой же цели может служить и подвижный маркер в виде креста, стрелки и т. д.

Вращение. Другим приемом, позволяющим изучать ДР в динамике, является получение последовательности ДР, полученных путем вращения трехмерного облака ОЕ вокруг некоторой оси, и изучения его двумерных проекций в фиксированном направлении. Таким образом, можно выбрать наиболее интересные двумерные проекции трехмерных точек. Итак, пусть имеем некоторое отображение наших ОЕ

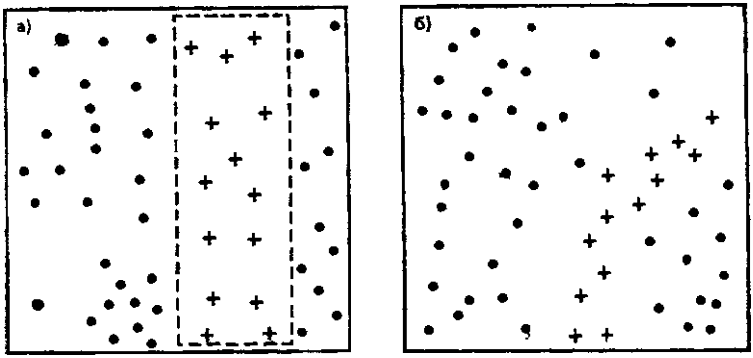


Рис. 18.2. Две проекции одного и того же множества объектов: а) пунктиром дано положение подвижного окна, крестами выделены точки, попавшие внутрь подвижного окна; б) положение тех же точек на другой проекции

в трехмерное пространство (например, пространство трех первых главных компонент или трех направлений целенаправленного проецирования по какому-либо из критериев и т. д.). Расположим оси $0x$ и $0y$ на экране дисплея, а ось $0z$ — перпендикулярно к нему. Начнем теперь вращать пространство вокруг оси $0x$ или $0y$, а направление проекции пусть остается ортогональным экрану. Выберем для определенности ось $0x$. Координаты ОЕ вдоль этой оси не меняются, а вертикальные координаты получаются из следующих уравнений:

$$\tilde{y}(t) = y\cos t + z\sin t \qquad (18.1)$$

или

$$\tilde{y}(t) = y(1-t) + zt. \qquad (18.2)$$

Если координаты y и z были нормированы, то новая координата $\tilde{y}$ также нормирована (имеет единственную дисперсию).

Обычно значение t берется с малым шагом и, если ЭВМ позволяет пересчитывать и подавать на экран дисплея ДР достаточно быстро, возникает плавная картина модификации изображения, своего рода фильм.

Вращение, задаваемое уравнением (18.1), отличается от задаваемого (18.2). Чтобы увидеть это, продифференцируем их по t. Имеем

$$\tilde{y}'(t) = z\cos t - y\sin t; \tag{18.3}$$

$$\tilde{y}'(t) = z - y. \tag{18.4}$$

Скорость изменения положения точек по вертикали для вращения (18.2) не зависит от t. В то же время для вращения (18.1) скорость изменяется с изменением угла вращения и в начале вращения скорость зависит только от неотображаемой визуально координаты z (это явление называется параллакс-эффектом [183]).

18.2.4. Обработка диаграмм рассеивания с помощью статистических методов. Рассмотренные ранее приемы манипуляции ДР, хотя и оказываются эффективными на практике, носят тем не менее технический характер. Способы обработки ДР, приведенные в настоящем параграфе, основаны на статистических идеях, и их целью является повышение «контрастности» структур, представленных на ДР, что позволяет легче обнаружить их визуально.

Рассматриваемый ниже подход основан на выделении k-ближайших соседей (см. гл. 7) для каждой ОЕ на ДР. При этом k-ближайших соседей выделяются либо в двумерном пространстве, соответствующем ДР, либо в исходном p-мерном пространстве. В качестве метрик может использоваться практически любая метрика, перечисленная в гл. 6, 11. Таким образом, данная процедура управляется тремя факторами: числом соседей; типом метрики; пространством переменных.

После выделения k-ближайших соседей получаем для каждой точки радиус минимальной сферы, в которую попали соответствующие k-соседей. Радиус такой сферы является монотонно убывающей функцией от оценки плотности распределения в данной точке по методу k-ближайших соседей. Теперь можно поступить по крайней мере двумя способами: 1) удалить заданный процент (5, 10, 20 %) точек с минимальной локальной плотностью; 2) позволить точкам дви-

гаться в направлении градиента оценки плотности (подробнее см. [233]).

Если на ДР есть какая-либо структура (например, кластерная), то обычно в результате одной из этих процедур она становится более выраженной визуально (см. [323]).

18.3. Преобразования данных в разведочном анализе данных

В данном параграфе речь идет о нелинейных преобразованиях исходных данных, представленных в виде матрицы «объект — признак». Нелинейные преобразования могут быть использованы в РАД: а) для линеаризации зависимостей между переменными, б) для упрощения структуры данных.

Линеаризация зависимостей между переменными. Цель использования таких преобразований состоит в переходе к новому набору переменных, зависимость между которыми является, возможно, более близкой к линейной. Если такое преобразование удается найти, то дальше к новой матрице данных можно с большим основанием применять такие линейные статистические методы, как главные компоненты, факторный анализ, линейную регрессию и т. д.

Будем рассматривать только преобразования вида

$$y^{(i)} = \varphi_i(x^{(i)}) \quad (i = \overline{1, p}),$$

где $\varphi_i(x^{(i)})$ — функции из некоторого класса допустимых функций Φ.

В качестве критерия, по которому ищется преобразование, можно использовать, например, критерий

$$Q = \sum_{\substack{i<j \\ i,\ j=1}}^{p} r_{ij}^2, \tag{18.5}$$

аналогичный критерию (17.30). Получить приближенное решение можно, если переменные $x^{(i)}$ предварительно градуировать (область значения переменной $x^{(i)}$ разбить на l_i градаций) и дальше использовать алгоритм из § 17.3.

Естественно, после градуирования для получения преобразований $\varphi_i(x_i)$ можно использовать и множественный анализ соответствий.

Дальше, в § 19.6, будет необходим случай максимизации (18.5), когда число переменных $p = 2$. Из регрессионного анализа известно [12, гл. 5], что, когда имеются две слу-

чайные величины $y^{(2)}$ и $x^{(1)}$, наилучшим, в смысле средней квадратической ошибки, регрессором вида $\varphi_1(x^{(1)})$ для случайной величины $y^{(2)}$ (т. е. для регрессии вида $y^{(2)} = \varphi_1(x^{(1)}) + \varepsilon$) будет условное математическое ожидание этой случайной величины при $x^{(1)}$, т. е. $\varphi_1(x^{(1)}) \sim \mathbf{E}(y^{(2)}/x^{(1)})$, и, следовательно, функция $\mathbf{E}(y^{(2)}/x^{(1)})$ имеет максимальный коэффициент корреляции с $y^{(2)}$. Аналогично верно и для регрессии $y^{(1)}$ на $x^{(2)}$. Поэтому функции $\varphi_2(x^{(2)})$ и $\varphi_1(x^{(1)})$ должны удовлетворять уравнениям

$$\begin{cases} \varphi_1(x^{(1)}) = c_1 \mathbf{E}(\varphi_2(x^{(2)})/x^{(1)}); \\ \varphi_2(x^{(2)}) = c_2 \mathbf{E}(\varphi_1(x^{(1)})/x^{(2)}). \end{cases} \tag{18.6}$$

Константы c_1 и c_2 не влияют на коэффициент корреляции. Кроме подхода, связанного с предварительным градуированием переменных, можно использовать и некоторые семейства монотонных преобразований, например преобразования Бокса — Кокса [196]:

$$\begin{cases} y^{(i)} = ((x^{(i)})^{\alpha_i} - 1)/\alpha_i, \ \alpha_i \neq 0; \\ y^{(i)} = \ln x^{(i)}, \ \alpha_i = 0 \end{cases} \tag{18.7}$$

или более обширное двухпараметрическое семейство

$$\begin{cases} y^{(i)} = (x^{(i)} - \beta_i)^{\alpha_i}/\alpha_i, \ \alpha_i \neq 0; \\ y^{(i)} = \ln(x^{(i)} - \beta_i), \ \alpha_i = 0. \end{cases} \tag{18.8}$$

Коэффициенты корреляции r_{ij} являются теперь функциями от α_k, β_k $(k = \overline{1, p})$ и задача (18.5) есть задача максимизации по этим параметрам.

Упрощение структуры данных. В этом случае стремятся получить преобразования, после применения которых распределение становится максимально похожим на многомерное нормальное. Используется некоторый класс преобразований, например (18.17), (18.8), но параметры α_i и β_i оцениваются уже не на основе максимизации критерия (18.5), а при максимизации функции правдоподобия.

Рассмотрим случай преобразования (18.7). Если предположим, что векторная случайная величина $Y = (y^{(1)}, \ldots y^{(p)})'$ подчинена многомерному нормальному распределению $N_p(\Theta, \Sigma)$, то для функции правдоподобия имеем следующее выражение:

$$p(X \mid \Theta, \Sigma, \alpha) = (2\pi)^{-n/2} |\Sigma|^{-n/2} J_\alpha \times$$

$$\times \exp\left[-\frac{1}{2} \sum_{i=1}^{n} (X_i^{(\alpha)} - \Theta')\Sigma^{-1}(X_i^{\alpha} - \Theta) \right], \tag{18.9}$$

где $\alpha = (\alpha_1, \ldots, \alpha_p)'$; $X_i^{(\alpha)} = \left(\left(x_i^{(1)}\right)^{\alpha_1}, \ldots, \left(x_i^{(p)}\right)^{\alpha_p}\right)$;

n — число объектов; $J_\alpha = \sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_i^{(j)}\right)^{\alpha_j - 1}$ —

якобиан преобразования.

Оценки параметров α_i ($i = \overline{1, p}$), Θ и Σ получаются из решения задачи

$$(\alpha, \Theta, \Sigma) \Rightarrow \underset{\widetilde{\theta}, \widetilde{\Sigma}, \widetilde{\alpha}}{\arg\max}\, p\left(X \mid \widetilde{\Theta}, \widetilde{\Sigma}, \widetilde{\alpha}\right),$$

Можно использовать и логарифм функции правдоподобия.

18.4. Использование дополнительных (иллюстративных) переменных и объектов

При использовании методов РАД существует опасность обнаружить в данных такие структуры, которые связаны, скорее, со спецификой данной выборки, но в силу ее недостаточного объема не отражают каких-либо устойчивых закономерностей в генеральной совокупности. В случае, когда исследуемое множество объектов само представляет собой всю генеральную совокупность, такой проблемы не возникает, однако если результаты, полученные при изучении выборки, будут использоваться для работы с объектами, не входящими в нее, проблема становится серьезной.

Конечно, можно надеяться, что этап «подтверждающего анализа» должен отсеять неправомерные со статистической точки зрения выводы. Однако некоторые возможности такого отсева существуют и в РАД. Один из приемов, применяемый с этой целью, состоит в разделении объектов и переменных на две части — активные (объекты, переменные) и иллюстративные, экзаменующие. Разделение объектов на «обучение» и «экзамен» широко используется в дискриминантном (см. гл. 3) и регрессионном анализе. Использование иллюстративных переменных менее распространено. Применение иллюстративных переменных в кластер-анализе описано в § 12.4.

Помимо проверки устойчивости выделенных структур, использование дополнительных элементов помогает и в интерпретации результатов РАД.

18.5. Основные типы данных и методы, используемые в разведочном анализе данных

РАД применяется к данным, заданным в одной из следующих форм:

матрица данных (МД) типа «объект — признак» с переменнымн, измеренными в количественных шкалах (МДК);

МД с переменными, измеренными в ординальной шкале (МДО);

МД с переменными, измеренными в номинальной шкале (МДН);

МД с переменнымн, измеренными в шкалах разной природы (гетерогенная МД, МДГ);

таблица данных типа «объект — объект» (будем рассматривать только случай матрицы расстояний (МР));

таблица сопряженностей (ТС).

Процедуры статистической обработки, используемые в РАД, могут быть разбиты на следующие группы в зависимости от целей анализа и типа обрабатываемых данных.

1. Вычисление основных статистических характеристик для матрицы типа МДК [10].
2. Преобразования переменных для МДК с целью линеаризации связей и (или) «нормализации» данных симметризации (см. § 18.3).
3. Преобразование переменных (оцифровка) для МДК, МДО, МДН, МДГ по различным критериям (см. § 17.4, 18.3).
4. Сокращение размерности данных с помощью линейных отображений: главные компоненты (ГК) (см. гл. 13), целенаправленное проецирование (гл. 19).
5. Нелинейные методы отображения данных типа МДК, МДО, МДН, МДГ (последние три матрицы в метрике Хемминга) (см. гл. 13).
6. Метрическое шкалирование для матриц типа МР (см. гл. 16).
7. Множественный анализ соответствий для МДО, МДН и МДГ, ТС (см. гл. 17).
8. Классификационные методы: кластер-анализ для таблиц МДК, МДО, МР, МДН, разделение смесей распределений, дискриминантный анализ (см. гл. 6—12).
9. Типологический анализ главных компонент (см. гл. 7). Анализ древообразной структуры зависимостей для МДК (см. гл. 2).

10. Кластер-анализ переменных МДК, МДО, МДН, МДГ, ТС. Пошаговый метод анализа структуры зависимостей переменных для МДО, МДН, ТС.
11. Анализ регрессионных зависимостей (метод целенаправленного проецирования, линейная модель) (см. гл. 19).

ВЫВОДЫ

1. Этап РАД применяется, когда у исследователя отсутствует априорная информация о статистическом или причинном механизме порождения имеющихся в его распоряжении данных. Основная цель РАД — построить некоторую статистическую модель данных (описания их структуры), которую, вообще говоря, необходимо дальше верифицировать. Можно сказать, что на этапе РАД формулируются статистические гипотезы, которые должны быть проверены на этапе подтверждающего анализа.
2. Важнейшим элементом РАД является широкое использование визуального представления многомерных данных, возможности которого возросли с появлением динамических форм визуального представления.
3. Преобразование данных в РАД позволяет либо линеаризовать связи между переменными, либо упростить для дальнейшего описание структуры данных.
4. Для верификации результатов РАД эффективным приемом является использование аппарата иллюстративных переменных и объектов.

Глава 19. ЦЕЛЕНАПРАВЛЕННОЕ ПРОЕЦИРОВАНИЕ МНОГОМЕРНЫХ ДАННЫХ

В этой главе в основном рассматриваются методы линейного проецирования данных. Совокупность таких методов в последнее время получила большое развитие и известна в зарубежной статистической литературе, как «projection pursuit» (PP).

Будем здесь использовать термин «целенаправленное проецирование» (ЦП). Методы ЦП являются естественным обобщением классических методов многомерного статистического анализа, таких, как факторный анализ, анализ глав-

ных компонент, линейный дискриминантный анализ и т. д. В отечественной литературе [36—40, 65, 67, 69, 104, 328] содержатся постановки ряда задач ЦП и методы их решения.

19.1. Цель и основные понятия целенаправленного проецирования

Метод ЦП [230, 246, 251, 328] основан на поиске наиболее «интересных» («выразительных») q-мерных линейных проекций исходных p-мерных данных $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$, где $q \ll p$. В РАД $q = 1, 2$, реже 3.

Пусть $\mathbf{U}$ — оператор линейного проецирования p-мерных данных на q-мерное пространство, т. е. набор из q *линейно независимых* p-мерных векторов $U_1, \ldots, U_q$, таких, что по определению $\mathbf{U}'X_k = (U_1' X_k, \ldots, U_q' X_k)'$, $1 \leqslant k \leqslant n$ и $Q(\mathbf{U}, X)$ — некоторая статистика, выборочное значение $\widetilde{Q}$ которой вычисляется по q-мерной выборке объема n. Тогда $\widetilde{Q} = Q(\mathbf{U}, \mathbf{X}^{(n)}) = \widetilde{Q}(U'X_1, \ldots, U'X_n)$ называется проекционным индексом (ПИ), характеризующим выразительность проекции $\mathbf{U}$ относительно статистики Q. Решение задачи РАД методом ЦП состоит из двух этапов:

1) выбор проекционного индекса $Q(\mathbf{U}, \mathbf{X}^{(n)})$;
2) поиск проекций $\mathbf{U}$, наиболее интересных относительно Q, т. е. решение задач:

найти

$$\mathbf{U} = \arg\max_{\widetilde{\mathbf{U}}} Q(\widetilde{\mathbf{U}}, \mathbf{X}^n). \tag{19.1}$$

Первому этапу посвящены следующие параграфы, здесь же кратко остановимся на втором. При решении задачи (19.1) для ряда важных ПИ $Q(\mathbf{U}, X^{(n)})$ удается использовать последовательный (пошаговый) метод получения проекционных векторов $U_1, \ldots, U_q$.

Допустим, что уже выбраны первые $(q-1)$ проекционных векторов $U_1, \ldots, U_{q-1}$. Тогда решается задача (19.1) в классе операторов $\mathbf{U} = (U_1, \ldots, U_q)$, где первые $(q-1)$ векторов — это отобранные ранее векторы, а U_q — любой линейно независимый с ними вектор. Иногда из формулы для $Q(\mathbf{U}, X^{(n)})$ ясно, что достаточно брать векторы $U_1, \ldots, U_q$ ортогональными, но в общем случае направления образуют косоугольную систему. Эта процедура может быть улучшена в результате использования дополнительного критерия «неинтересности» направления проецирования. Тогда в алгоритм можно включать шаги, на которых «неинтересные» направления выбрасываются. В каждом из рассмотренных ва-

риантов пошаговый метод реализуется обычными процедурами *условной* оптимизации (условия линейной независимости, ортогональности или S-ортогональности, где S — например, ковариационная матрица). Имеются важные ПИ, для которых пошаговый метод не эффективен. В этом случае необходимо вернуться к оптимизационной задаче (19.1) в исходной постановке, т. е. решать ее как задачу безусловной оптимизации на многообразии всех операторов q-мерного проецирования. Численные процедуры решения таких задач разработаны в [37—39] и рассмотрены в гл. 20.

Прежде чем перейти к последующему изложению, кратко остановимся на вопросе, почему собственно используются линейные отображения? Имеется несколько обоснований различной природы для использования линейных отображений многомерных данных для целей анализа. Перечислим некоторые из них (оговорим, что порядок перечисления не отражает их относительной важности).

Во-первых, линейные отображения приводят к тому, что в качестве новых переменных в пространстве образов используются линейные комбинации исходных переменных. Это существенно упрощает интерпретацию выделяемых структур (например, кластеров), поскольку позволяет использовать такие хорошо освоенные в статистике понятия, как факторные нагрузки или вклады переменных (нормированные тем или иным способом коэффициенты линейных комбинаций).

Во-вторых, имеется важное статистическое обоснование, связанное со статистическими свойствами линейных проекций многомерных случайных величин. Именно при достаточно широких предположениях относительно плотности распределения многомерной случайной величины X [215] распределение случайно выбранной линейной комбинации переменных стремится к нормальному, когда $p \to \infty$. На практике это означает, что при достаточно большом числе переменных подавляющее большинство линейных комбинаций исходных переменных будет иметь «почти» нормальное распределение.

Поскольку нормальное распределение является некоторым эталоном распределения, не обладающего какой-либо из перечисленных в § 18.1 структур (за исключением структуры типа эллипсоидального рассеивания), при поиске этих структур можно выбирать линейные комбинации, распределение которых наиболее сильно отличается от нормального. В частности, в качестве ПИ можно использовать любые критериальные величины, применяемые для проверки гипотезы нормальности.

В-третьих, имеется довольно общая статистическая модель для кластерной структуры в виде смеси эллипсоидально симметричных распределений, рассматриваемая в следующем параграфе. Оказывается, что вся информация о кластерах содержится в некотором линейном подпространстве R^{+}, называемом дискриминантом подпространства. Если компонентами смеси будут нормальные распределения, то снова придем к разложению исходного пространства на два компонента — «интересный», имеющий распределение, отличное от нормального, и содержащий линейные комбинации с нормальным распределением.

19.2. Проекционные индексы, подходящие для выделения кластеров

19.2.1. Смеси эллипсондально снмметрнчных распределений как модель кластерной структуры. Будем предполагать, что плотность распределения $p(X)$, генерирующего выборку $\mathbf{X}^{(n)}$, представляет собой смесь унимодальных эллиптическн симметричных плотностей

$$p(X) = \sum_{i=1}^{k} a_i d_i(X), \tag{19.2}$$

где

$$d_i(X) = c(d, p, \mathbf{W})\, d((X - M_i)' \mathbf{W}^{-1} (X - M_i)); \tag{19.2'}$$

$c(d, p, \mathbf{W})$ — нормирующая константа; $a_i > 0, \sum_{i=1}^{k} a_i = 1$ — веса компонента смеси; $d(y)$ — некоторая неотрицательная, монотонно убывающая при $y \to \infty$ функция $(y^{p-1} d(y)\, dy < \infty)$; M_i — вектор средних i-й компоненты смеси; $\mathbf{W}$ — невырожденная матрица ковариаций (внутрикомпонеитного рассеивания), одинаковая для всех компонент.

В частности, если $d(y) = \exp(-y/2)$, то $d_i(X)$ будет плотностью нормального распределения. (Некоторые другие примеры плотностей приведены в § 20.1.)

Смесь плотностей вида (19.2), (19.2') можно рассматривать как одну из возможных моделей для описания кластерной структуры. Плотность $p(X)$ имеет k модальных значений (если компоненты смеси достаточно разнесены), и точки в окрестностн какой-либо модальной точки можно считать относящимися к одному и тому же кластеру.

Матрицу ковариаций для случайного вектора с плотностью $p(X)$ можно представить в виде $\mathbf{S} = \mathbf{B} + \mathbf{W}$, где $\mathbf{B}$— матрица межкомпонентного рассеивания

$$\mathbf{B} = \sum_{i=1}^{k} a_i (M_i - M_0)(M_i - M_0)';$$

$M_0 = \sum_{i=1}^{k} a_i M_i$ — вектор средних значений для X. Дальше, не ограничивая общности, для простоты будем считать, что величина X центрирована, т. е. $M_0 = 0$.

Пусть теперь $z = U'X$ — некоторая одномерная проекция. Плотность случайной величины z есть k-компонентная смесь симметричных унимодальных распределений

$$f(z) = \sum_{i=1}^{k} a_i e_i(z), \quad e_i(z) = e((z - m_i)/w)/w,$$

где $e(z) = c(d, p, I_p) \int d\left(z^2 + \sum_{i=1}^{k-1} y_i^2\right) dy_1 \ldots dy_{k-1}$;

$m_i = U' M_i$; $w^2 = U' \mathbf{W} U$.

Дисперсия z равна $s^2 = b^2 + w^2$, где $b^2 \quad b^2(U)$ — величина межкомпонентного разброса для z, т. е. $b^2 = U' \mathbf{B} U$.

Введем отношение

$$t^2(U) = b^2(U)/w^2(U), \tag{19.3}$$

которое можно рассматривать как меру различия компонент смеси для одномерной проекции, задаваемой вектором U.

Поиск направлений проецирования, максимизирующих отношение $t^2(U)$, приводит к *каноническим переменным*.

19.2.2. **Дискриминантное подпространство.** В дискриминантном анализе используются так называемые канонические переменные $v^{(i)} = V_i' X$ ($i = 1, \ldots, q^+$) (см. [129]), где векторы V_i ($i = 1, \ldots, q^+$) суть собственные векторы с положительными собственными значениями $t_1, \ldots, t_{q^+} > 0$ задачи $(\mathbf{B} - t\mathbf{W}) V = 0$. Число $q^+ \leqslant \min(p, k - 1)$ и зависит от геометрической конфигурации векторов средних M_i ($i = \overline{1, k}$). В частности, если центры компонент смеси лежат на одной прямой, то $q^+ = 1$. Векторы V_i будут $\mathbf{B}$-ортогональными, $\mathbf{W}$—ортогональными, и, следовательно, $\mathbf{S}$-ортогональными. Величина собственного числа t_i равна значению $t^2(V_i)$, т. е. отношения (19.3) для направления проецирования V_i.

Подпространство $R^+ = \text{span}(V_1, \ldots, V_{q^+})$ называется *дискриминантным подпространством* (ДП) и содержит пол-

ную информацию о различиях среди компонент смеси (19.2), другое эквивалентное определение этого подпространства будет: $R^+ = \text{span}\,(\mathbf{W}^{-1} M_1, \ldots, \mathbf{W}^{-1} M_k)$.

В связи с вышесказанным следует, что проекционные векторы для ЦП (в рамках модели (19.2), (19.2′)) должны принадлежать R^+.

Оценка ДП является одной из задач дискриминантного анализа. Однако в ДА считается, что известны или могут быть оценены обе матрицы **B** и **W**. Оцеика матрицы **W** производится по обучающим выборкам (ОВ), т. е. в дискриминантном анализе матрица $\mathbf{X}^{(n)}$ должна быть разбита на k подматриц $\mathbf{X}_i^{(n_i)}$ $(i = \overline{1, k})$ относительно объектов (наблюдений), из которых известно, что они принадлежат i-й компоненте смеси (19.2).

Если же ОВ нет, то может быть оценена только матрица и приходится использовать другие подходы.

19.2.3. Проекционные индексы, использующие математическое ожидание монотонных функций плотности одномерной проекции. Рассмотрим однопараметрическое семейство проекционных индексов (ПИ) для одномерных проекций, задаваемых вектором U.

$$Q_\beta(U, X) = s^\beta \mathbf{E}_f f^\beta(z), \quad (\beta > 0), \tag{19.4}$$

где $\mathbf{E}_f$ — оператор математического ожидания по плотности $f(z)$.

Приведем без доказательств неравенства, связывающие значение $Q_\beta(U, X)$ и отношение $t^2(U) = b^2/w^2$ в рамках модели (19.2):

$$g(e, \beta)\left(\sum_{i=1}^{k} a_i^{\beta+1}\right)(1 + t^2(U))^{\beta/2} \leqslant Q_\beta(U, X) \leqslant$$

$$\leqslant g(e, \beta)(1 + t^2(U))^{\beta/2}, \tag{19.5}$$

где константа $g(e, \beta) = \mathbf{E}_e e^\beta(z)$ не зависит от U.

В частности, если имелась смесь нормальных распределений, то $g(e, \beta) = 1/\left((\sqrt{2\lambda})^\beta \sqrt{1+\beta}\right)$.

Можно показать, что когда $t^2(U) = 0$, то $Q_\beta(U, X) = g(e, \beta)$, т.е. точной будет правая граница. Величина $g(e, \beta)$ является минимальной, достигаемой индексом $Q_\beta(U, X)$. С другой стороны, левая граница асимптотически достигается, если все попарные расстояния Махаланобиса $t^2_{ij} = (m_i - m_j)^2/w^2$ между компонентами смеси неограниченно возрастают, т. е. $t^2_{ij} \to \infty$. Поэтому можно ожидать, что если имеются проекции, где компоненты смеси хорошо разделены, то они будут найдены решением соответ-

ствующей (19.4) максимизационной задачи. Конечно, это, скорее, эвристическое соображение, нежели точные рассуждения (можно, в частности, показать, что $Q_\beta(U, X)$ не является монотонной функцией $t^2(U)$).

Пример 19.1. Приведем выражение для вычисления $Q_\beta(U, X)$ в случае смеси нормальных распределений при $\beta = 1$.

$$Q_1(U, X) = g(e, 1)(1 + t^2(U))^{1/2}\left(1 + 2\sum_{i>j}^{k} a_i a_j e^{-t_{ij}^2/4}\right).$$

Для нормальной плотности величина $g(e, 1) = 1/(2\sqrt{\pi})$.

Когда $\beta \to 0$, критерий (19.4) переходит в энтропийный критерий

$$Q_0(U, X) = -\int f(z) \ln (sf(z))\, dz. \qquad (19.6)$$

Все приведенные выше эвристические соображения могут быть применимы и к (19.6).

Замечание. Можно использовать и отрицательные значения β в (19.4). Тогда, однако, нужно либо искать направления U, минимизирующие величину $Q_\beta(U, X)$ ($\beta < 0$), либо переходить к ПИ вида $1/Q_\beta(U, X)$ или $-Q_\beta(U, X)$ и снова решать для последних задачу на максимальное значение.

19.2.4. Проекционные индексы, основанные на использовании моментов третьего и четвертого порядков. Идея использования момента третьего порядка для поиска направлений, хорошо отображающих кластеры (если они есть), достаточно очевидна, если предполагать верной модель смеси симметричных распределений. Пусть U — проекционный вектор, тогда третий момент для одномерной проекции запишется

$$\mu_3(U) = \sum_{i=1}^{k} a_i m_i^3. \qquad (19.7)$$

Дальше всюду, без ограничения общности, будем считать данные центрированными, т. е. полагать $\sum_{i=1}^{k} a_i M_i = 0$, тогда $\sum_{i=1}^{k} a_i m_i = 0$.

Из выражения (19.7) видно, что отличие $\mu_3(U)$ от нуля обусловлено только несовпадением средних значений компонент смеси ($m_1 = M_1' U$). Конечно, даже при несовпадении средних μ_3 может быть равен 0 для любой проекции, напри-

мер, для любой проекции двухкомпонентной смеси с равными весами $a_1 = a_2 = 1/2$.

В качестве ПИ в решении максимизационной задачи целесообразнее использовать не сам третий момент, а коэффициент асимметрии $\gamma_1 = \mu_3^2/s^6$. Хотя возможно и непосредственное использование $\mu_3(U)$ для восстановления дискриминантного подпространства (см. п. 19.3.2).

Использование четвертого момента и связанного с ним коэффициента эксцесса γ_2 как ПИ основано на том соображении, что если имеется смесь нормальных плотностей, проекциям, на которых компоненты смеси не разделены, соответствует нулевое значение коэффициента эксцесса γ_2. Для выделения выразительных проекций, вообще говоря, следует решать две задачи — искать как проекции, доставляющие максимум γ_2, так и проекции, доставляющие минимум. Выражение для четвертого момента одномерной проекции имеет вид:

$$\mu_4(U) = \sum_{i=1}^{k} a_i (M_i' U)^4 - 3\left(\sum_{i=1}^{k} a_i (M_i' U)^2\right)^2 + \tilde{c}(U' S U)^2. \quad (19.8)$$

Константа $\tilde{c}$ зависит только от функции $d(y)$. В частности, для нормального распределения $\tilde{c} = 3$. Коэффициент эксцесса тесно связан с ПИ, предложенными Краскалом в работе [259].

$$v_{\text{Краск}} = \sqrt{\mathrm{D}d^{\alpha}}/\mathrm{E}d^{\alpha},$$

где $\mathrm{D}d^{\alpha}$ — дисперсия расстояний в степени α между точками из $\mathbf{X}^{(n)}$; $\mathrm{E}d^{\alpha}$ — среднее значение α-х степеней расстояний Т. е. $v_{\text{Краск}}$ — это коэффициент вариации α-х степеней расстояний.

Дж. В. Краскал предлагал использовать значения $\alpha < 1$, в частности $\alpha = 2/7$. Однако, как показывает опыт практического использования таких ПИ, на самом деле более эффективно использовать $\alpha > 1$. Легко показать, что при $\alpha = 2$ $v_{\text{Краск}} = \gamma_2 + 1$. Как и при использовании моментов третьего порядка, для восстановления дискриминантного подпространства не обязательно решать оптимизационную задачу с γ_2. Альтернативный подход используется в п. 19.3.2.

19.2.5. Проекционные индексы, основанные на распределении разностных векторов. В задачах кластерного анализа и разделения смесей важной характеристикой структуры данных является распределение разностного вектора $X_{i1} - X_{i2}$.

Предположим, что p-мерный случайный вектор X имеет плотность распределения $p(X)$. Введем ПИ

$$Q_{2,\alpha}(U, X) = \frac{1}{2\alpha} P\left\{\frac{|z_1 - z_2|}{2} \leqslant \alpha s\right\}, \qquad (19.9)$$

где $z_i = U'X_i$, $\alpha > 0$. Плотность распределения случайной величины $z = z_1 - z_2$ имеет вид:

$$g_U(z) = \int_{-\infty}^{+\infty} f_U(z_1) f_U(z_1 + z)\, dz_1.$$

Следовательно,

$$Q_{2,\alpha}(U, X) = s \int_{-\infty}^{+\infty}\int f_U(z_1) f_U(z_1 + st)\, \chi(t, \alpha)\, dt dz_1,$$

где $\chi(t, \alpha)$ — плотность равномерного распределения на интервале $|-\alpha, \alpha|$. Таким образом, $Q_{2,\alpha}(U, X)$ является оценкой $Q_1(U, X)$ и при $\alpha \to 0$ $Q_{2,\alpha}(U, X) \to Q_1(U, X)$ ($Q_1(U, X) = Q_\beta(U, X)$ при $\beta = 1$);

В случае, когда имеется матрица данных $\mathbf{X}^{(n)}$, в качестве оценки значения ПИ $Q_{2,\alpha}$ естественно взять частоту встречаемости пар векторов $X_{i1} - X_{i2}$, таких, что $|U'X_{i1} - U'X_{i2}| < \alpha s$, где s — выборочное стандартное отклонение.

Обратим внимание, что указанную оценку для ПИ $Q_{2,\alpha}$ можно использовать для поиска и q-мерных выразительных проекций $q > 1$.

Выборочное значение ПИ $Q_{2,\alpha}$ дает оценку значения ПИ Q_1 (19.4) и тем самым еще одну содержательную интерпретацию ПИ $Q_\beta(U, X)$. Покажем, что соответствующая интерпретация ПИ $Q_\beta(U, X)$ существует и для всех *целых* β.

Пусть $X_1, \ldots, X_n$ — выборка из генеральной совокупности случайного вектора X с плотностью $p(X)$. Введем проекционный индекс

$$Q_{r,\alpha}(U, X) = \frac{1}{c_r \alpha^{r-1}} P\{S(Z^{(r)}) \leqslant \alpha s\},$$

где r — произвольное целое;

$Z^{(r)} = (U'X_1, \ldots, U'X_r)'$ и $S^2(Z^{(r)}) = \frac{1}{r} \sum |z_j - z_0|^2$,

$z_0 = \frac{1}{r}\sum_{j=1}^{r} z_j$; c_r — объем единичного шара в r — 1-мерном пространстве. Плотность распределения r-мерной случайной величины $(z_1 - z_0, \ldots, z_r - z_0)$ сосредоточена на под-

пространстве, задаваемом уравнением $z_1 + \ldots + z_r = 0$ и имеет вид:

$$g_U(z_1, \ldots, z_r) = \int f_U(z_1 + t) \ldots f_U(z_r + t)\, dt,\ z_1 + \ldots + z_r = 0.$$

Следовательно,

$$Q_{r,\alpha}(U, X) = s^{r-1} \int \ldots \int f_U(z_1 + t) \ldots f_U(z_r + t)\, \chi_\alpha(z^{(r)})\, dt dz^{(r)},$$

где $\chi_\alpha(z^{(r)})$ — плотность равномерного распределения в шаре с центром 0, радиуса α в подпространстве $\sum_{j=1}^{r} z_j = 0$. В случае, когда задана матрица $\mathbf{X}^{(n)}$, выборочной оценкой значения ПИ является частота встречаемостн в данной выборке объема n подвыборок объема r, стандартное отклонение которых не превосходит αs, где s — стандартное отклонение всей выборки. Те же рассуждения, что и выше, показывают, что выборочное значение ПИ $Q_{r,\alpha}$ дает оценку значения ПИ $Q_{\beta=r-1}(U, X)$.

19.3. Выявление эллипсоидальной кластерной структуры (восстановление дискриминантного подпространства)

19.3.1. Восстановление дискриминантного подпространства на основе проекционных индексов типа функционалов от плотностей распределения проекций. Почему решение оптимизационной задачи (19.1) с использованием ПИ вида (19.4) приведет к выявлению кластерной структуры? Частично ответ следует из рассмотрения неравенства (19.5). Более того, оказывается, что, решая пошаговым методом задачу (19.1), придем к некоторому новому базису в ДП, которое, как указывается в п. 19.2.2, содержит полную информацию о кластерной структуре в случае, если верна модель (19.2), (19.2′). Верна следующая лемма.

Л е м м а 19.1. Пусть имеет место модель смеси распределений (19.2), (19.2′). Предположим теперь, что векторы U_1, ..., U_{q^+} найдены с помощью последовательной (step-wise) процедуры максимизации ПИ $Q_\beta(U, X)$ (19.4) и при этом каждый из векторов U_i S-ортогонален подпространству, натянутому на векторы U_1, ..., U_{i-1}, т. е. к span $(U_1, \ldots, U_{i-1})$. Тогда каждый вектор U_i принадлежит к дискриминантному подпространству R^+ и, более того, $R^+ = \text{span}(U_1, \ldots, U_{q^+})$. Таким образом, векторы U_1, ..., U_{q^+} образуют некоторый базис в R^+, отличный от канонического V_1, ..., V_{q^+}.

Доказательство. ПИ $Q_\beta(U, X)$ есть некоторая функция от G от $s, w, m_1, \ldots, m_k$, т. е. $Q_\beta(U, X) = (G(s, w, m_1, \ldots, m_k)$. Каждая же из величин $s, w, m_1, \ldots, m_k$ есть функция вектора U. Дифференцируя G по U и приравнивая производную нулю, получим уравнение, которому необходимо должен удовлетворять вектор, максимизирующий $Q_\beta(U, X)$:

$$(\dot{G}_s/s)\,\mathbf{S}U + (\dot{G}_w/w)\,\mathbf{W}U + \sum_{i=1}^{k} \dot{G}_{m_i} M_i = 0,$$

где через $\dot{G}_1$, $\dot{G}_w$, $\dot{C}_{m_i}$ обозначены соответствующие частные производные. Умножим это уравнение слева на $\mathbf{W}^{-1}$, что дает после некоторых преобразований

$$(h(U)\,\mathbf{I}_p + \mathbf{W}^{-1}\mathbf{B})\,U = V,$$

где

$$V = \sum_{i=1}^{k} \dot{G}_{m_i}\,\mathbf{W}^{-1} M_i, \quad h(U) = \dot{G}_w + \dot{G}_s. \tag{19.10}$$

Вектор V является линейной комбинацией векторов $\mathbf{W}^{-1} M_i$, каждый из которых принадлежит R^+ (предполагаем X центрированным) и, следовательно, сам вектор $V \in R^+$. Теперь покажем, что вектор U_1, максимизирующий $Q_\beta(U, X)$, принадлежит R^+. Предположим, что это не так, т. е. $U_1 = c_1 U^+ + c_2 U^-$, где $U^+ \in R^+$ и $U^- \in R^-$ (R^- есть S-ортогональное дополнение к R^+) и $c_2 \neq 0$. Подстановка U_1 в (19.10) приводит к следующему уравнению:

$$c_2 h(U_1) U^- = V - c_1 (h(U_1)\,\mathbf{I}_p + \mathbf{W}^{-1}\mathbf{B})\,U^+. \tag{19.10'}$$

Вектор в правой части этого равенства принадлежит R^+. С другой стороны, значение $h(U) = 0$ только, если $t^2(U) = 0$, т. е. если $U_1 \in R^-$ (это можно проверить непосредственным вычислением производных G_s' и G_w'). Следовательно, равенство (19.10′) верно, если $U_1 \in R^-$ (и тогда $V = 0$, $c_1 = 0$ и U_1 не является максимизирующим вектором) или если $U^- = 0$. Итак, вектор U_1, максимизирующий $Q_\beta(U, X)$, принадлежит к R^+. Аналогично доказывается, что векторы $U_2, \ldots, U_{q^+}$ также принадлежат R^+ (при условии попарной S-ортогональности). Так как rank $(R^+) = q^+$ и эти векторы S-ортогональны, то $\mathbf{R}^+ = \text{span}\,(U_1, \ldots, U_{q^+})$.

Заметим теперь, что число q^+ обычно неизвестно. Однако и в этом случае лемма 19.1 позволяет получить некоторые полезные следствия. Например, если $q^+ = 3$, то первые три вектора U_1, U_2, U_3 позволяют извлечь всю информацию о различиях между компонентами смеси. Когда же

$q^+ > 3$, но собственные числа $t_4, \ldots, t_{q^+}$ достаточно малы по сравнению с t_3, то по соображениям непрерывности те же самые три вектора U_1, U_2, U_3 извлекают главную часть такой информации. С другой стороны, если все собственные числа t_i примерно одинаковы, безразлично, какие векторы брать для проецирования, лишь бы они принадлежали R^+, но это обеспечивается.

Другими словами, если задача поддается визуализации, т. е. имеется проекция размерности $q \leqslant 3$, на которой компоненты смеси хорошо разделены, то она будет получена с помощью критерия (19.4).

Приведем пример, показывающий, что условие равенства внутрикомпонентных матриц ковариаций является существенным для оценки дискриминантного подпространства на основе максимизации ПИ типа $Q_\beta(U, X)$.

Пример 19.2. Рассмотрим двумерное распределение

$$p(X) = \frac{1}{2}(N(M_1, W_1) + N(M_2, W_2)),$$

где $M_1' = \left(0, -\frac{m}{\sqrt{1+m^2}}\right)$, $M_2' = \left(0, \frac{m}{\sqrt{1+m^2}}\right)$,

$$\mathbf{W}_1 = \begin{pmatrix} \sqrt{2-1/\alpha} & 0 \\ 0 & 1/\sqrt{1+m^2} \end{pmatrix},$$

$$\mathbf{W}_2 = \begin{pmatrix} \sqrt{1/\alpha} & 0 \\ 0 & 1/\sqrt{1+m^2} \end{pmatrix}.$$

Легко проверить, что X имеет среднее, равное нулю, и единичную ковариационную матрицу. Проекция X на первую координатную ось имеет плотность вида

$$f_1(z) = \frac{1}{2}(N(0, \sqrt{2-1/\alpha}) + N(0, \sqrt{1/\alpha})),$$

а на вторую ось

$$f_2(z) = \frac{1}{2}(N(-m/\sqrt{1+m^2}, 1/\sqrt{1+m^2})).$$

Заметим, что $f_1(z)$ не зависит от m, а $f_2(z)$ не зависит от α, поэтому

1) для любого α существует такое $m_0 = m_0(\alpha)$, что для всех $m > m_0$ критерий Q_1 (19.4) достигает максимума на проекции $U = (1, 0)'$;

2) для любого m существует такое $\alpha_0 = \alpha_0(m)$, что для всех $\alpha \geqslant \alpha_0$ критерий Q_1 достигает максимума на проекции $U = (0, 1)'$.

В обоих случаях для выделения кластеров, скажем, по критерию дискриминантного анализа или визуально, предпочтительнее вторая координатная ось. На этой оси расстояние махаланобисского типа между компонентами смеси максимально.

В то же время следует отметить, что проекция на первую координатную ось обладает следующим экстремальным свойством: различие по вторым моментам (отношение дисперсии первого компонента смеси ($\sqrt{2}$ при $\alpha \to \infty$) к дисперсии второго компонента (0 при $\alpha \to \infty$) максимально). Таким образом, можно сказать, что в условиях, когда модель смеси с равными ковариационными матрицами неверна, проекции, получаемые из условия максимума критерия, могут быть экстремальными как в отношении неоднородности средних значений компонент, так и в отношении неоднородности дисперсий.

19.3.2. Оценка дискриминантного подпространства на основе моментных индексов. Использование ЦП на основе критериев вида (19.4) на практике требует значительного объема вычислений. В данном параграфе предлагается простой способ оценки ДП или нескольких векторов на него на основе критериев асимметрии и эксцесса [69].

Полученные таким способом направления проецирования могут использоваться как самостоятельно, так и как «хорошие» стартовые точки для определения направлений проецирования на основе критерия (19.4).

Рассмотрим способ получения векторов, математическое ожидание которых принадлежит не самому ДП, а подпространству $R_M^{\perp} = \operatorname{span}(M_1, \ldots, M_k)$. Переход к ДП осуществляется умножением этих векторов на матрицу S^{-1}. (Напомним, что величина X предполагается центрированной).

Используя (19.7), (19.8), докажем следующую лемму.

Лемма 19.2. Пусть по выборке получены векторы

$$\widehat{U}_1(U) = (1/n) \sum_{j=1}^{n} (U' X_j)^3 X_j, \tag{19.11}$$

$$\widehat{U}_2(U) = (1/n) \sum_{j=1}^{n} (U' X_j)^3 X_j - \widetilde{cs}^2 SU, \tag{19.12}$$

где S — оценка матрицы ковариаций X; U — произвольный вектор.

Тогда для математических ожиданий векторов (19.11), (19.12) верны соотношения

$$\begin{cases} \mathbf{E}\widehat{U}_1(U) = c_1(n) \sum_{i=1}^{k} a_i (M_i' U)^2 M_i; & (19.13) \\ \mathbf{E}\widehat{U}_2(U) = c_2(n) \left(\sum_{i=1}^{k} a_i (M_i' U)^3 M_i - \right. & \\ \left. - \tilde{c} \left(\sum_{i=1}^{k} a_i m_i^2 \right) \sum_{i=1}^{k} a_i m_i M_i \right), & (19.13') \end{cases}$$

где константы $c_1(n)$ и $c_2(n)$ зависят только от n и $c_1(n), c_2(n) \to 1$, когда $n \to \infty$.

Из (19.13), (19.13') следует, что

$$\mathbf{E}\widehat{U}_1(U) \in R_M^+ \text{ и } \mathbf{E}\widehat{U}_2(U) \in R_M^+.$$

Докажем только равенство (19.13) для вектора (19.11). Имеем

$$\mathbf{E}\left[(1/n) \sum_{j=1}^{n} (U' X_j)^3 \right] = c_1(n) \mu_3(U)$$

(в левой части в квадратных скобках стоит смещенная оценка третьего момента μ_3). Возьмем производную по U от обеих частей этого равенства. Так как дифференцирование — линейная операция, то в левой части равенства его можно провести под знаком оператора усреднения $\mathbf{E}$, что и дает соотношения (19.13).

Следующая лемма определяет способ получения оценки некоторого вектора из R_M^+, свободный от произвола в выборе вектора U.

Лемма 19.3. Пусть $\mathbf{H}$ есть некоторая положительно определенная матрица, а

$$U(\mathbf{H}) = \sum_{i=1}^{k} a_i (M_i' \mathbf{H} M_i) M_i. \quad (19.14)$$

Тогда для вектора

$$\widehat{U}(\mathbf{H}) = (1/n) \sum_{j=1}^{n} (X_j' \mathbf{H} X_j) X_j \quad (19.15)$$

верно равенство $\mathbf{E}\widehat{U}(\mathbf{H}) = c_1(n)\ U(\mathbf{H})$. Действительно, пусть вектор U в (19.11) распределен независимо от X с средним 0 и матрицей ковариаций $\mathbf{H}$. И пусть $\mathbf{E}_U$ — оператор усреднения по U. Тогда $\mathbf{E}_U(\mathbf{E}\widehat{U}_1(U)) = \mathbf{E}_U U_1 = U(\mathbf{H})$.

Операторы $\mathbf{E}_U$ и $\mathbf{E}$ в левой части равенства можно поменять местами, поскольку U и X независимы. Но $\mathbf{E}_U\ \widehat{U}_1(U) = \widehat{U}(\mathbf{H})$, что и доказывает лемму 19.3.

19.3.3. Оценка подпространства R_M^+. Пусть теперь $U_0, \ldots, U_{q^+-1}$ — последовательность векторов вида $U_{j+1} = \sum_{c=1}^{k} a_i \times (U'_j M_i)^2 M_i$.

Вектор U_0 задается формулой (19.14). Каждый из векторов $U_i \in R_M$. Предположим, что ранг набора векторов $U_0, \ldots, U_{q^+}$ равен q^+, тогда верна следующая лемма (дается без доказательства).

Лемма 19.4. Пусть последовательность векторов $\widehat{U}_j$ задается соотношением

$$\widehat{U}_{j+1} = (1/n) \sum_{l=1}^{n} (\widehat{U}'_j X_l) X_l,$$

где $\widehat{U}_0$ определяется из выражения (19.15). Тогда

$$\mathbf{E}\|\widehat{U}_j - U_j\|^2 \sim 0(1/n),\ j = 0, \ldots, q^+ - 1.$$

Поскольку ранг системы векторов $U_0, \ldots, U_{q^+-1}$ равен q^+, то подпространство $\widehat{R}_M^+$, натянутое на векторы $\widehat{U}_0, \ldots, \widehat{U}_{q^+-1}$, будет являться оценкой для R_M^+.

Хотя в реальной ситуации ранг q^+ неизвестен, можно все же построить оценку R_M^+, например, с помощью следующей процедуры.

Пусть $R_{\widehat{U}}^i$ — подпространство, натянутое на $\widehat{U}_0, \ldots, \widehat{U}_{i-1}$, а η_i — угол между $R_{\widehat{U}}^i$ и $\widehat{U}_{i+1}$. Можно показать, что углы векторов $\widehat{U}_{q^+}, \ldots, \widehat{U}_p$ с $R_{\widehat{U}}^{q^+}$ и тем более углы $\widehat{\eta}_{q^+}, \ldots, \widehat{\eta}_p$ должны стремиться к 0. Анализируя последовательность углов η_i, можно определить номер $\widehat{q^+}$, начиная с которого они становятся малы, и в качестве оценки для R_M^+ взять $R_{\widehat{U}}^{\widehat{q^+}}$.

19.4. Проекционные индексы для дискриминантного анализа

Как направления проецирования в ДА можно использовать канонические направления по Рао. Таким образом, в качестве ПИ выступает отношение t^2 (19.3). В случае двух классов придем к единственному направлению — дискриминантной функции Фишера (см. п.1.1.2). Однако использование канонических направлений эффективно только тогда, когда соответствующая структура может быть описана смесью вида (19.2), (19.2′) с равными матрицами внутрикомпонентного рассеивания и, что, пожалуй, самое главное, расстояния Махаланобиса между классами должны быть достаточно велики. Кроме того, оценка матрицы ковариаций **W** и средних чувствительны к наличию аномальных наблюдений.

Предлагаемые в п. 19.4.1, 19.4.2 подходы позволяют иногда построить направления проецирования, которые дают картину взаимного расположения объектов из разных классов в ситуациях, отличающихся от модели (19.2), (19.2′).

19.4.1. Проекционные индексы для линейных классификаторов. Пусть p-мерная выборка $\mathbf{X}$ разбита на две подвыборки $\mathbf{X}_1 = \{\mathbf{X}_{1,1}, \ldots, \mathbf{X}_{n_1,1}\}$ и $\mathbf{X}_2 = \{X_{1,2}, \ldots, X_{n_2,2}\}$. В рамках классической модели ДА (построение линейного классификатора) наиболее интересной одномерной проекцией этой выборки является решение задачи ЦП для ПИ:

$$Q(U) = \frac{|U'\overline{X}_1 - U'\overline{X}_2|}{s_U}, \tag{19.16}$$

где $\overline{X}_k$ — средний вектор выборки $\mathbf{X}_k$, $k = 1, 2$ и s_U — среднеквадратическое отклонение проекции выборки $\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2$. В качестве робастного варианта такого ПИ рассматривается

$$Q(U) = \frac{\operatorname{med}(U'\mathbf{X}_1) - \operatorname{med}(U'\mathbf{X}_2)}{\operatorname{mad}(U'\mathbf{X})}. \tag{19.17}$$

Здесь med — медиана, а mad — *медиана абсолютных отклонений*, например, mad $U'\mathbf{X}$ — медиана последовательности $|y_{i_1} - y_{i_2}|$, где y_{i_1}, y_{i_2} пробегают выборку $U'\mathbf{X}$. В [246] П. Хьюбер особо рекомендует следующую модификацию ПИ:

$$Q(U) = \frac{\operatorname{med}(U'\mathbf{X}_1) - \operatorname{med}(U'\mathbf{X}_2)}{\operatorname{mad}\{(U'\mathbf{X}_1 - \operatorname{med} U'\mathbf{X}_1) \cup (U'\mathbf{X}_2 - \operatorname{med} U'\mathbf{X}_2)\}}. \tag{19.18}$$

В тех случаях, когда нет оснований для классической модели ДА даже в робастном вариаите, желательно использовать проекционные индексы, опирающиеся на более детальную информацию о распределении разностного вектора $X_1 - X_2$, $X_k \in \mathbf{X}_k$, $k = 1, 2$.

Рассмотрим проекционный индекс $Q_\lambda(U) = \frac{1}{2\lambda} \times$ $\times P(|U'(X_1 - X_2)| \leqslant \lambda)$, где λ — задаваемый, априорный порог разрешимости и $||U|| = 1$. Он относится к тем ПИ, для которых критерий выразительности непосредственно заложен в их построение.

Пусть $f_k(X)$ — плотность распределения случайного p-мерного вектора X_k и $f_k(y, U)$ — индуцированная плотность распределения проекции $y_k = U'X_k$. Тогда проекция разностного вектора $U'(X_1 - X_2)$ имеет плотность распределения $f(y, U) = \int\limits_{-\infty}^{\infty} f_1(y_1, U) f_2(y + y_1, U)\, \mathrm{d}\, y_1$ и поэтому можно записать $Q_\lambda(U) = \int\limits_{-\infty}^{\infty} \chi_1(y, \lambda) f(y, U)\, \mathrm{d}y = \iint\limits_{-\infty}^{\infty} \chi_1 \times$ $\times (y_1 - y_2, \lambda) f_1(y_1, U) f_2(y_2, U)\, \mathrm{d}y_1\, \mathrm{d}y_2$, где $\chi_1(y, \lambda)$ — плотность равномерного распределения на отрезке $[-\lambda_1, \lambda]$. Таким образом, в теоретическом случае при малых λ ПИ $Q_\lambda(U)$ близок к ПИ:

$$Q_0(U) = \int\limits_{-\infty}^{\infty} f_1(y, U) f_2(y, U)\, \mathrm{d}y.$$

Сравним выборочные варианты этих ПИ. Пусть, как и выше, заданы две обучающие выборки $\mathbf{X}_1$ и $\mathbf{X}_2$.

Тогда в качестве $\widehat{Q}_\lambda(U)$ — выборочного варианта ПИ $Q_\lambda(U)$ — возьмем

$$\widehat{Q}_\lambda(U) = \frac{1}{2\lambda} \widehat{P}(|U'(X_{i,1} - X_{j,2})| \leqslant \lambda),\ 1 \leqslant i \leqslant n_1,\ 1 \leqslant j \leqslant n_2),$$

где $\widehat{P}(\cdot)$ — частота, а $\widehat{Q}_0(U)$ построим следующим образом:

выберем оценку плотности $f_k(y, U)$ в виде $\widehat{f}_k(y, U) = \frac{1}{n_k} \sum\limits_{i=1}^{n_k} \chi_1\left(y - y_{i,k}, \frac{\lambda}{2}\right)$, где $y_{i,k} = U'X_{i,k}$. Тогда

$$\widehat{Q}_0(U) = \int\limits_{-\infty}^{\infty} \widehat{f}_1(y, U) \widehat{f}_2(y, U)\, \mathrm{d}y = \frac{1}{n_1 n_2} \times$$

$$\times \frac{1}{\lambda^2} \sum_{i,j} (\lambda - |y_{i,1} - y_{j,2}|)_+. \qquad (19.19)$$

Здесь $z_+ = \begin{cases} 0, & z < 0 \\ z, & z \geqslant 0 \end{cases}$. Заметим, что $\frac{d}{dz} z_+ = z_+^0 = \begin{cases} 0, & z < 0 \\ 1, & z \geqslant 0 \end{cases}$, поэтому

$$\widehat{Q}_\lambda(U) = \frac{1}{n_1 n_2} \frac{1}{2\lambda} \sum_{i,j} (\lambda - |y_{i,1} - y_{j,2}|)_+^0. \tag{19.19'}$$

Сравнивая формулы (19.19) и (19.19'), приходим к следующему результату:

$$\widehat{Q}_\lambda(U) = \widehat{Q}_0(U) + \frac{\lambda}{2} \frac{d}{d\lambda} \widehat{Q}_0(U). \tag{19.20}$$

Докажем формулу (19.20). Для любых i, j и $\lambda > 0$ непосредственное вычисление показывает, что

$$\frac{(\lambda - |y_{i,1} - y_{j,2}|)_+^0}{2\lambda} = \frac{(\lambda - |y_{i,1} - y_{j,2}|)_+}{\lambda^2} +$$

$$+ \frac{\lambda}{2} \frac{d}{d\lambda} \frac{(\lambda - |y_{i,1} - y_{j,2}|)_+}{\lambda_2}. \tag{19.21}$$

Разделив (19.21) на $n_1 n_2$ и просуммировав по i, j, получаем формулу (19.20).

В многоклассовой задаче, когда $\mathbf{X} = \bigcup_{s=1}^{k} \mathbf{X}_s$, где $\mathbf{X}_s$—s-я p-мерная обучающая выборка объема n_s; обозначим через $\mathbf{X}_{s,t}$ массив разностных векторов $\{\mathbf{X}_{i,s} - \mathbf{X}_{j,t}\}$. Положим

$$\widehat{Q}_{\lambda;\, s,\, t}(U) = \frac{1}{2\lambda} \widehat{P}\{|U'Z| \leqslant \lambda,\ Z \in \mathbf{X}_{s,t},\ s < t\}; \tag{19.22}$$

$$\widehat{Q}_\lambda^k(U) = \frac{1}{2\lambda} \widehat{P}\{|U'Z| \leqslant \lambda,\ Z \in \bigcup \mathbf{X}_{s,t},\ 1 \leqslant s < t \leqslant k\}. \tag{19.23}$$

Ясно, что

$$\widehat{Q}_\lambda^k(U) = \sum_{s,t} \pi_{st} \widehat{Q}_{\lambda,\, st}(U), \tag{19.24}$$

где $\pi_{st} = \frac{n_s n_t}{\Sigma n_s, n_{t'}}$. Таким образом, ПИ (19.24) является скаляризацией матрицы критериев $\widehat{Q}_{\lambda,\, st}(U)$. На основе скаляризации этой матрицы строятся и другие ПИ, например,

$$\widehat{Q}_\lambda(U) = \max \widehat{Q}_{\lambda,\, st}(U); \tag{19.25}$$

$$\widehat{Q}_\lambda(U) = \Sigma \pi_{st} \alpha_{st} \widehat{Q}_{\lambda,\, st}(U), \tag{19.26}$$

где $\alpha_{st} \geqslant 0$ — матрица штрафов за ошибки неправильной классификации. Отметим еще один способ построения ПИ в многоклассовой задаче. Образуем массив $\mathbf{V} = \{V = (X_1, \ldots, X_k)\}$, где $X_k \in \mathbf{X}_k$, т. е. $\mathbf{V}$ — массив наборов V представителей классов, и положим

$$\widehat{Q}_\lambda^{\mathbf{V}}(U) = \frac{1}{c_{k-1}\lambda^{k-1}} \widehat{P}\{s_k(U'V) \leqslant \lambda^2\}, \tag{19.27}$$

где $s_k(U'V) = \frac{1}{k}\sum_{s=1}^{k} |U'X_s - U'\bar{V}|^2$, $\bar{V} = \frac{1}{k}\sum_{s=1}^{k} X$ и c_{k-1} — объем единичного шара в $(k-1)$-мерном пространстве R^{k-1}.

Для $p \times k$-мерного случайного вектора $V = (X_1, \ldots, X_k)$ его k-мерная проекция $(y_1 - \bar{y}, \ldots, y_k - \bar{y})$, где $y_s = U'X_s$, а $\bar{y} = \frac{1}{k}\sum_{s=1}^{k} y_s$ имеет плотность распределения

$$F(y_1, \ldots, y_k)\big|_{\Sigma y_s = 0} = \int_{-\infty}^{\infty} f_1\left(y_1 + \frac{t}{k}, U\right) \ldots$$

$$\ldots f_k\left(y_k + \frac{t}{k}, U\right) dt,$$

поэтому ПИ (19.27) является оценкой теоретического проекционного индекса

$$Q_\lambda^{\mathbf{V}}(U) = \int_{-\infty}^{\infty} \ldots \int \chi_{k-1}(y_1 - \bar{y}, \ldots, y_{k-1} - \bar{y}, \lambda) f_1(y_1, U) \ldots$$

$$\ldots f_k(y_k, U)\, dy_1 \ldots dy_k,$$

где $\chi_{k-1}(\cdot, \lambda)$ — плотность равномерного распределения на $(k-1)$-мерном шаре радиуса λ. Таким образом, в качестве $Q_\lambda^{\mathbf{V}}(U)$ естественно взять

$$Q_0^{\mathbf{V}}(U) = \int_{-\infty}^{\infty} f_1(y, U) \ldots f_k(y, U)\, dy. \tag{19.28}$$

Выбирая, как и выше, оценку $\widehat{f}_s(y, U)$ плотности $f_s(y, U)$, после несложных вычислений получаем:

$$\widehat{Q}_0^{\mathbf{V}}(U) = \int_{-\infty}^{\infty} \widehat{f}_1(y, U) \ldots \widehat{f}_k(y, U)\, dy = \frac{1}{n_1 \ldots n_k} \times$$

$$\times \frac{1}{\lambda^k} \sum_{(y_{i_1, 1}, \ldots, y_{i_k, k})} (\lambda - \max(y_{i_1, 1}, \ldots, y_{i_k, k}) -$$

$$- \min(y_{i_1, 1}, \ldots, y_{i_k, k})_+. \tag{19.28'}$$

Для ПИ $\widehat{Q}_0^{\mathbf{V}}(U)$ имеется аналог соотношения (19.20).

Обозначим через $V_{i_1, \ldots i_n}(U)$ набор $(y_{i_1, 1}, \ldots, y_{i_l, k})$, где $y_{i_s, s} = U' X_{i_s, s}$. Тогда (19.28′) можно переписать в виде

$$\widehat{Q}_0^{\mathbf{v}}(U) = \frac{1}{n_1 \ldots n_k} \frac{1}{\lambda^k} \sum_{i_1 \ldots i_k} (\lambda - w(V_{i_1 \ldots i_k}(U)))_+,$$

где $w(V_{i_1 \ldots i_k}(U))$— размах набора $(y_{i_1, 1}, \ldots, y_{i_k, k})$ представителей выборок $\mathbf{X}_1, \ldots, \mathbf{X}_k$. Всего таких наборов, очевидно, $n_1 \ldots n_k$.

Имеем

$$\frac{(\lambda - w)_+^q}{\lambda^{k-1}} = k \frac{(\lambda - w)_+}{\lambda^k} + \lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{(\lambda - w)_+}{\lambda^k}.$$

Следовательно, ПИ $\widehat{Q}_0^{\mathbf{v}}(U)$ связан с ПИ

$$\widehat{Q}_\lambda^*(U) = \frac{1}{k\lambda^{k-1}} \widehat{P}\{w(y_{i_1, 1}, \ldots, y_{i_k, k}) \leqslant \lambda\} \tag{19.29}$$

соотношением

$$\widehat{Q}_\lambda^*(U) = \widehat{Q}_0^{\mathbf{v}}(U) + \frac{\lambda}{k} \frac{\mathrm{d}}{\mathrm{d}\lambda} \widehat{Q}_0^{\mathbf{v}}(U). \tag{19.30}$$

Проекционные индексы (19.19), (19.19′), (19.24) хорошо зарекомендовали себя при решении задач технической и медицинской диагностики (распознавании образов) и используются с начала 70-х годов [38, 39, 70, 104].

Для поиска «выразительной» проекции $\mathbf{U}: R^p \to R^q$, $\mathbf{U} = (U_1, \ldots, U_q)$, доставляющей минимум этим ПИ, в [104] был применен пошаговый алгоритм условной оптимизации, в котором после того, как найдены векторы $U_1, \ldots, U_n$, $n < q$, следующий вектор U_{n+1} ищут как решение задачи:

$$U_{n+1}^* = \arg\min_{U_{n+1}} \widehat{P}\left\{\sum_{l=1}^{n+1} |U_l' Z|^2 \leqslant d^2 : \|U_{n+1}\| = 1,\right.$$

$$\left. U_{n+1} \perp U_1 \perp \ldots \perp U_n;\ B_{n+1}\right\},$$

где $\perp$ — символ ортогональности, Z — разностный вектор, а условие B_{n+1} означает, что в построении очередного вектора U_{n+1} участвуют только те разностные векторы Z, длина проекции которых на подпространство с базисом $U_1, \ldots, U_n$ меньше d. Когда объемы $n_1, \ldots, n_k$ выборок $\mathbf{X}_1, \ldots, \mathbf{X}_k$ ве-

лики, алгоритм применяется к выборкам их типичных представителей, полученным предварительно, например, при помощи процедур автоматической классификации. В этом случае часто удается получить результат при помощи ПИ:

$$Q(U) = \max \| Z - \mathbf{U}' \mathbf{U} Z \|^2, \qquad (19.31)$$

где Z пробегает разностные векторы типичных представителей.

Алгоритмы поиска выразительных проекций, реализующие методы безусловной оптимизации сразу на всем многообразии всех ортогональных проекций из R^p в R^q, разработаны в [37—39]. В [38] дано детальное описание алгоритма минимизации ПИ (19.31), основанного на методе градиентного спуска в задаче векторной оптимизации.

19.4.2. Проекционные индексы и направления в задаче классификации нормальных распределений с неравными ковариационными матрицами. Здесь рассматривается случай $k = 2$ классов. В этом случае, если матрицы ковариаций классов равны, существует единственное направление проецирования (размерность q^+ для ДП R^+ равна 1). И это направление есть *дискриминантный вектор Фишера* (см. гл. 1). В принятых здесь обозначениях

$$U_1 = \mathbf{W}^{-1}(M_2 - M_1). \qquad (19.32)$$

В случае, когда матрицы внутриклассового рассеивания не равны ($\mathbf{W}_1 \neq \mathbf{W}_2$), направление (19.32) можно получить, используя матрицу $\mathbf{W} = a_1 \mathbf{W}_1 + a_2 \mathbf{W}_2$. Однако в этой ситуации возможно построить и другие направления проецирования. Более того, можно получить направления проецирования и для случая, когда $M_1 = M_2$ (центры групп совпадают).

Один из способов получения вектора U_2 предложен в [301]. В качестве U_2 используется вектор, получаемый из условия максимума ПИ

$$Q(V, X) = \frac{(V', M_2 - M_1)^2}{V' \mathbf{W} V} \qquad (19.33)$$

при дополнительном условии ортогональности $(U_1' U_2) = 0$, т. е.

$$U_2 = \underset{V,\ (V' U_1) = 0}{\operatorname{argmax}} Q(V, X).$$

В результате получается следующее выражение для U_2:

$$U_2 = \left(\mathbf{W}^{-1} \Delta + \frac{\Delta' \mathbf{W}^{-2} \Delta}{\Delta' \mathbf{W}^{-1} \Delta} \mathbf{W}^{-2} \Delta \right), \quad \Delta = M_2 - M_1. \qquad (19.34)$$

Недостаток этого подхода состоит в том, что вектор U_2 определен и тогда, когда $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}$, хотя для нормальных распределений в этом случае имеется только одно направление проецирования — вектор Фишера.

Еще один подход, отличный от предлагаемого далее для построения векторов $U_2, \ldots, U_q$, дополнительных к вектору Фишера, дан в работе [101].

Рассмотрим процедуру построения проекционных векторов для ПИ, зависящих от моментов первого и второго порядка для первого и второго классов (так как нормальные распределения отличаются только по этим характеристикам). Ограничимся построением только одного вектора U_2. Более полное изложение дано в [67].

Меру расстояния для одномерных распределеиий, соответствующих проекциям компонент G_1 и G_2 на вектор V и зависящую от первых двух моментов, можно записать в виде $R^2(U) = R^2(\omega_1^2(U), \omega_2^2(U), \Delta^2(U))$,
где $\Delta^2(U) = (m_1 - m_2)^2$.
В качестве $R^2_{(U)}$ можно выбрать расстояние Махаланобиса, дивергенцию Кульбака [91], расстояние Бхаттачария и др. (см. гл. 1). Для того чтобы построить ПИ, введем понятие условного расстояния и среднего условиого расстояния.

Условное расстояние между проекциями компонент (классов) на вектор V, когда проекция точки X на некоторый другой вектор U равна z, $z = U'X$ определяется как расстояние между *соответствующими условными нормальными распределениями* с параметрами $m_i(V/z)$, $\omega_i^2(V/z)$. Заметим, что дисперсии $\omega_i^2(V/z)$ не зависят от конкретиого значения z, а зависят только от направления U [67], т. е. можно записать: $\omega_i^2(V/z) = \omega_i^2(V/U)$. В то же время величина $\Delta(V, z) = m_2(V/z) - m_1(V/z)$ есть линейная функция z.

Дадим теперь определение *среднего условного расстояния* между проекциями компонент на вектор V:

$$R^2(V/U) = \mathrm{E}R^2(V/z) = \int R^2(V/z) \sum_{i=1}^{2} a_i \varphi(z;$$

$$m_i(U), \omega_i^2(U))\,\mathrm{d}z,$$

где $R^2(V/z) = R^2(\omega_1^2(V/U), \omega_2^2(V/U), \Delta^2(V/z))$;

$\varphi(z; m, \omega^2)$ — плотность нормального распределения с параметрами m и ω^2.
Величиина $R^2(V/U)$ и является проекционным иидексом.

Пусть в качестве вектора U_1 выбираем вектор Фишера (19.32) (это только один из возможных вариантов). Тогда,

если в качестве расстояния использовать величину (19.34), в качестве вектора U_2, максимизирующего (19.34) (соответствующее этой величине аналитическое выражение приведено в [67]), получим векторы

$$\widetilde{U}_2 = \mathbf{W}^{-1}(\mathbf{W}_1 - \mathbf{W}_2)\, U_1; \tag{19.35}$$

$$U_2 = ORT_{U_1}(\widetilde{U}_2),$$

где $ORT_{U_1}(\widetilde{U}_2)$ — составляющая вектора U_2, ортогональная U_1.

19.5. Выделение аномальных наблюдений

19.5.1. Проекционный индекс и приближенная вычислительная процедура. В качестве ПИ, подходящего для получения проекций, на которых аномальные наблюдения (outliers) могли бы наблюдаться визуально, можно воспользоваться отношением

$$Q(U,\ \mathbf{X}^{(n)}) = s^2(U)/s^2_{\text{уст}}(U), \tag{19.36}$$

где $s^2(U)$ — обычная оценка дисперсии одномерной проекции выборки $\mathbf{X}^{(n)}$ на вектор U; $s^2_{\text{уст}}(U)$ — некоторая устойчивая оценка параметра масштаба.

Известно, что обычная оценка $s^2(U)$ весьма чувствительна к наличию аномальных наблюдений и их присутствие приводит, как правило, к возрастанию ее величины. Поэтому те направления, на которых значения ПИ (19.36) достигают максимума, могут обоснованно рассматриваться как направления, где влияние аномальных наблюдений наиболее выражено (если, конечно, таковые вообще имеют место).

В числителе (19.36) стоит квадратичная форма $s^2(U) = U'SU$, знаменатель приближенно можно аппроксимировать квадратичной формой $s^2_{\text{уст}} \cong U'\mathbf{S}_{\text{уст}}U$, где $\mathbf{S}_{\text{уст}}$ — некоторая устойчивая оценка матрицы ковариаций. Поэтому как приближенное решение оптимизационной задачи для (19.36) можно использовать решение обобщенной задачи на собственные значения и векторы

$$(\mathbf{S} - h\mathbf{S}_{\text{уст}})\, U = 0. \tag{19.37}$$

Имеется не более p положительных собственных чисел для задачи (19.37), которые можно упорядочить в порядке убывания их величины $h_1 \geqslant h_2 \geqslant \ldots \geqslant h_q > 1$. Для получения проекций используются собственные векторы $U_1, \ldots, U_q$, соответствующие наибольшим собственным числам, превосходящим 1.

Устойчивые оценки матрицы ковариаций и вектора средних. Устойчивые оцеики матрицы ковариаций можно получать разными методами. В частностн, имеющаяся в пакете ППСА [66] программная реализацня основана на использованни разновидности M-оценок [269], так называемых экспоненциально-взвешенных оценок [11, гл. 10]. Однако экспоненцнально-взвешенные оценкн обладают тем недостатком, что в случае дискретных переменных с некоторым значением, частота которого больше частот остальных значений (что часто встречается на практике), оценкой матриц коварнаций может быть матрица с нулями на диагоналн, т. е. оценки дисперсий для этих переменных равны нулю, что иногда приводит к трудностям в реализации процедуры.

Модификация индекса выразительности (19.36). Критерий (19.36) можно усовершенствовать, если учесть еще различие между оцеиками параметров положеиия (обычиой M и устойчивой $M_{\text{уст}}$), например, положив

$$Q(U,\ \mathbf{X}^{(n)}) = (s^2(U) + \| m - m_{\text{уст}} \|^2)/s^2_{\text{уст}}(U),$$

где $m = M'U$, $m_{\text{уст}} = M'_{\text{уст}}U$.

Приближенное решение снова получается как решение полной проблемы собственных векторов и чисел

$$(\mathbf{S} + (M - M_{\text{уст}})(M - M_{\text{уст}})' - h\mathbf{S}_{\text{уст}}) = 0.$$

Пример 19.3. Рассмотрим пример применения метода главных компонент и ЦП к выборке реальиых данных. Используем матрицу данных из работы [149], содержащую сведеиия о 130 сельскохозяйственных районах СССР за 1975 г. Показатели, использованные в этой матрице, представляют собой некоторые обобщенные характеристики: возрастной состав населения, состав сельскохозяйственной продукции, техническую оснащенность и т. д. Всего имеется 26 таких показателей ($p = 26$), каждый из них имеет пять градаций, измерены они в ординальиой шкале.

Результаты примеиеиия метода главиых компонент в ЦП приведены соответственно на рис. (19.1, *а*, *б*) , где квадратами обозначены 5% наблюдений, имеющих мипимальиый вес $\omega_i = (X_i - M_{\text{уст}})'\mathbf{S}^{-1}_{\text{уст}}(X_i - M_{\text{уст}})$ (они рассматриваются в качестве «подозрительных» как аномальиые наблюдеиия). На рис. (19.1, *а*) эти паблюдения хорошо выделепы и далеко отстоят от осиовиой массы наблюдений.

Важно, однако, знать, действительно ли эти наблюдения могут в каком-либо содержательном смысле играть роль аномальных? Идентификация этих паблюдений показывает, что им соответствуют Магаданская, Архангельская, Мур-

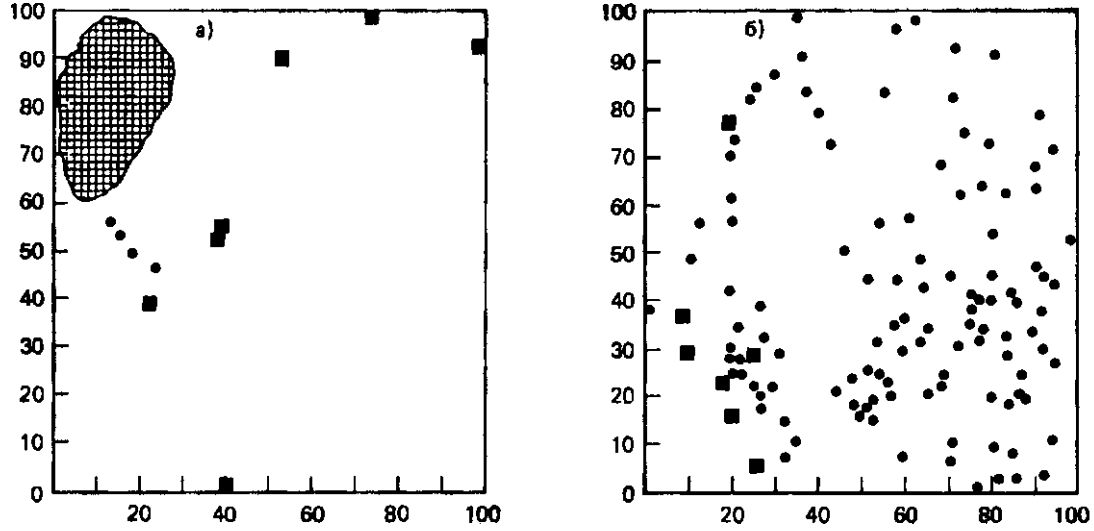


Рис. 19.1. Диаграмма рассеивания для 130 сельскохозяйственных регионов СССР: а) целенаправленное проецирование для выделения аномальных наблюдений, точки, «подозрительные» как аномальные, обозначены закрашенными квадратами; б) отображение тех же объектов на плоскости двух первых главных компонент

маиская и т. д. области. В смысле структуры сельскохозяйственного производства это действительно районы, резко отличающиеся от большинства сельскохозяйственных районов СССР — сельское хозяйство в них направлено в основном на удовлетворение нужд крупного промышленного города (Магадана, Архангельска и т.д.) и почти ннчего не производит для других потребнтелей в СССР.

19.6. Выделение нелинейных структур в многомерных данных

Значительный нитерес при анализе многомерных данных вызывает наличие в нпх нелинейных структур, т. е. концентрации распределения в окрестности некоторого нелинейного многообразия размерности $q \ll p$.

Разумеется, столь же интересно наличне н линейных многообразий, в окрестности которых концентрируется распределение. Однако линейные многообразия достаточно хорошо могут быть выделены с помощью, например, метода главных компонент. Здесь же рассмотрим применение ЦП для выделения нелинейных многообразий.

В качестве ПИ может быть нспользован любой критерий независимости. Действительно, пусть $U_1, \ldots, U_q$ — базис пространства отображения, причем векторы U_i $(i = \overline{1,q})$ выбраны так, чтобы случайные величины $z^{(i)} = (U_i'X)$ были линейно независимы (нескоррелированы), т. е. cov $(z^{(i)}, z^{(j)}) = 0$, $i \neq j$. Для этого необходимо и достаточно, чтобы векторы U_i были попарно S-ортогональнымн, поскольку cov $(z^{(i)}, z^{(j)}) = U_i' SU$. Тогда наличне какой-лнбо структуры в пространстве отображения означает, что переменные $z^{(1)}, \ldots, z^{(q)}$ должны быть зависимы. При этом, поскольку исключили линейную зависимость между переменнымн $z^{(1)}, \ldots, z^{(q)}$, эта структура не может быть описана с помощью линейных фуикций от них.

При выборе критернев независимости, подходящих в качестве ПИ, нужно учитывать еще следующие факторы: возможность получения выборочной оценки критерия, простой в вычислительном отношении (нбо именно она будет иа практике нспользоваться в качестве ПИ), и возможность быстрой оценки граднента ПИ.

Предлагаемые ниже ПИ основаны на использовании определения независимости набора случайных величин [11]: случайные величины $z^{(1)}, \ldots, z^{(q)}$ распределены независимо тогда и только тогда, когда их совместная функция распре-

деления может быть представлена в виде произведения маргинальных функций распределения

$$(F(t_1, \ldots, t_q) = P(z^{(1)} < t_1, \ldots, z^{(q)} < t_q) = \prod_{i=1}^{q} F_i(t_i), \quad (19.38)$$

где $F_i(t_i) = P(z^{(i)} < t_i)$ — маргинальная функция распределения для $z^{(i)}$.

Из (19.38) можно получить аналогичные соотношения для плотностей и т. д.

Перейдем теперь к формулировке ПИ.

19.6.1. Интегральное квадратичное расхождение. Для непрерывных случайных величин в качестве ПИ можно использовать следующую величину:

$$Q(U, X) = |\Sigma_Z|^{1/2} \int \left(p(Z) - \prod_{i=1}^{q} f_i(z^{(i)}) \right)^2 dZ, \quad (19.39)$$

где $f_i(z^{(i)})$ — плотность распределения одномерной проекции $z^{(i)} = U_i'X$; $p(Z)$ — плотность совместного распределения; Σ_Z — матрица ковариаций для Z, диагональная в силу выбора U_i.

ПИ (19.39) инвариантен относительно преобразований масштаба в пространстве Z и аффинноинвариантен относительно преобразований в пространстве X. Однако относительно вращений в пространстве Z этот ПИ неинвариантен, поскольку при этом могут меняться маргинальные функции плотности $f_i(z^{(i)})$.

Можно получить некоторую «разумную» аффинноинвариантную относительно преобразований Z разновидность ПИ (19.39), заменив маргинальные плотности $f_i(z^{(i)})$ плотностями нормального распределения. При этом, учитывая инвариантность (19.39) относительно линейных преобразований X, можно заранее перейти в пространстве X к махаланобисовой метрике (см. § 5.2). Векторы $U_1, \ldots, U_q$ будем выбирать ортонормированными. Соответствующий критерий будет иметь вид

$$Q(U, X) = \int \left(p(Z) - \prod_{i=1}^{q} \varphi(z^{(i)};\ 0,\ 1) \right)^2 dZ, \quad (19.40)$$

где $\varphi(z;\ 0,1)$ — плотность стандартного нормального распределения.

Этот критерий направлен на поиск q-мерных проекций, индуцированное распределение для которых наиболее сильно отличается от стандартного q-мерного нормального распределения с независимыми компонентами.

Поскольку, как указывалось в § 19.1, известно, что невыразительные проекции имеют, при широких предположениях, нормальное распределение, критерий (19.40) будет обладать достаточной общностью. Для поиска одномерных проекций такой критерий предлагается в [65].

Нормальное распределение с независимыми компонентами в (19.40) выступает, таким образом, в качестве эталона бесструктурности.

Возможна дальнейшая полезная модификация критерия (19.40) на основе следующего приема. Если случайные величины z распределены по закону $N_q(0, I_q)$, то случайные величины

$$y^{(i)} = \Phi(z^{(i)}), \tag{19.41}$$

где $\Phi(z)$ — функция нормального стандартного распределения, распределены равномерно в единичном кубе с вершинами (0, ..., 0), (1, 0, ..., 0) и т. д. Интеграл (19.40) после преобразования (19.41) переходит (с точностью до множителя, не зависящего от неизвестных векторов U_1, ..., U_q) в

$$\tilde{Q}(U, X) = \int_{\square} (\tilde{p}(Y) - 1)^2 \, dY, \tag{19.42}$$

где $\tilde{p}(Y)$ — плотность распределения, а область интегрирования — единичный куб.

Здесь в качестве эталона однородности выступает *равномерное распределение в единичном кубе.*

Элементарное преобразование (19.42) приводит снова к критерию типа среднего значения степени плотности

$$\tilde{Q}(U, X) = \int_{\square} \tilde{p}^2(Y) \, dY, \tag{19.42'}$$

поскольку проекции, максимизирующие (19.42) и (19.42'), совпадают. Критерий, аналогичный (19.42), предложен в [227].

19.6.2. «Наивные» ПИ на основе параметризации вида зависимости. Хотя сами случайные величины $z^{(1)}, \ldots, z^{(q)}$ линейно независимы (т. е. $\operatorname{cov}(z^{(i)}, z^{(j)}) = 0$ при $i \neq j$), можно попытаться установить наличие зависимости между ними, используя некоторые функции от них и изучая линейную зависимость между этими функциями.

Пусть $q = 2$. Будем искать функции $\psi_1(z^{(1)})$ и $\psi_2(z^{(2)})$, такие, чтобы коэффициент корреляции между $\psi_1(z^{(1)})$ и $\psi_2(z^{(2)})$ был бы максимальным. Решение этой задачи дано в § 18.3.

Однако, если ограничиться конкретным классом функций, например полиномов от $z^{(i)}$, можно получить решение задачи максимизации коэффициента корреляции в аналитическом виде, что, конечно, существенно удобнее для реализации вычислительных процедур по максимизации критерия.

В частности, ограничиваясь двумя степенями от $z^{(i)}$, можно использовать такие ПИ:

$$\int Q_{\mathrm{I}}(U, X) = r^2(z^{(1)}, y_2) + r^2(y_1, z^{(2)}); \qquad (19.43)$$

$$| Q_{\mathrm{II}}(U, X) = Q_{\mathrm{I}}(U, X) + r^2(y_1, y_2), \qquad (19.43')$$

где $y_i = (z^{(i)})^2$, $r(\cdot, \cdot)$ — соответствующий коэффициент корреляции.

Приведем аналитическое выражение как функцию компонент векторов проецирования, например для $r(z^{(1)}, y_2)$. Для упрощения обозначений положим, что $z^{(1)} = (U'X)$, $z^{(2)} = (V'X)$. Кроме того, будем считать, что $(U'V) = 0$, $\mathbf{E}X = 0$, $\mathbf{S} = \mathbf{I}_q$ (махаланобисова метрика). Тогда $\mathbf{E}z^{(1)} = \mathbf{E}z^{(2)} = 0$, $\mathbf{D}z^{(1)} = \mathbf{D}z^{(2)} = 1$, $\mathbf{E}y_1 = \mathbf{E}y_2 = 1$, $\mathbf{D}y_2 = \mathbf{E}(y_2 - 1)^2 = \mathbf{E}y_2^2 - 1$.
Далее

$$\mathbf{E}y_2^2 = \mathbf{E}(V'X)^4 = \sum_{i_1, i_2, i_3, i_4 = 1}^{p} v_{i_1} v_{i_2} v_{i_3} v_{i_4} \times$$

$$\times \mathbf{E}[x^{(i_1)} x^{(i_2)} x^{(i_3)} x^{(i_4)}]. \qquad (19.44)$$

Коэффициент корреляции $r(z^{(1)}, y_2) = \mathbf{E}z_1(y_2 - 1)/\sqrt{\mathbf{D}y_2}$. Отсюда получаем

$$r(z^{(1)}, y_2) = \sum_{i_1, i_2, i_3 = 1}^{p} v_{i_1} u_{i_2} u_{i_3} \mathbf{E}[x^{(i_1)} x^{(i_2)} x^{(i_3)}]/\sqrt{\mathbf{E}y_2^2 - 1}, \qquad (19.45)$$

где значение $\mathbf{E}y_2^2$ определяется формулой (19.44). Аналогичные формулы получаются и для остальных коэффициентов корреляции. Это дифференцируемые функции от компонент U и V. Для вычисления производных нужно знать значения смешанных третьих и четвертых моментов компонент вектора X (на практике используются их оценки)

19.7. Регрессия на основе целенаправленного проецирования

Пошаговая аддитивная процедура аппроксимации функции регрессии. Подход для аппроксимации функции регрессии с использованием ЦП предложен в работе [229]. Пусть име-

ется выборка объема n из $(p+1)$-мерного распределения вектора Y и необходимо восстановить функцию регрессии $y^{(p+1)}$-й компоненты на p первых компонент вектора. Далее для упрощения формул будем употреблять обозначение y вместо $y^{(p+1)}$ и X для вектора, составленного из p первых компонент вектора Y. Предположим, что функцию регрессии можно представить в виде

$$y = \sum_{i=1}^{q} g_i(U_i X) + \varepsilon, \tag{19.46}$$

где $g_i(\cdot)$ — неизвестные функции; U_i — неизвестные векторы; q — число проекций, которое также неизвестно.

Уравнение (19.46) может рассматриваться как развитие обобщенной линейной модели [12].

Вычислительная процедура состоит в следующем.

На первом шаге ищут такую функцию $g_1(\cdot)$ и вектор U_1, чтобы

$$\delta_1^2 = \sum_{j=1}^{n} (y_j - g_1(U_1' X_j))^2 \Rightarrow \min. \tag{19.47}$$

Этот поиск осуществляется следующим образом. Задавая некоторую проекцию U_1, ищут непараметрическую оценку функции $g_1(\cdot)$, например с помощью сплайн-аппроксимации, минимизирующую δ_1^2. Далее при фиксированной функции $g_1(\cdot)$ ищут новый вектор U_1. Затем снова настраивается функция $g_1(\cdot)$ и т.д. до тех пор, пока значение $\delta_1^2(U_1, g_1)$ не стабилизируется. После этого от величин y_j переходят к остаткам $\tilde{y}_j = y_j - g_1(U_1, X_j)$.

Поиск вектора U_2 и функции $g_2(\cdot)$ проводится теперь из условия минимизации величины

$$\delta_2^2 = \sum_{j=1}^{n} (\tilde{y}_j - g_2(U_2 X_j))^2$$

описанным выше способом.

Данный процесс итерируется до тех пор, пока остаточная сумма квадратов δ_q^2 для некоторого q не станет меньше порогового значения. Доказано [631], что регрессия в форме (19.46) точно восстанавливает истинную функцию регрессии, если последняя имеет вид полинома некоторой степени от компонент X. В качестве примера в [229] рассмотрен случай, когда $p = 2$ и истинная зависимость между y и $x^{(1)}$ и $x^{(2)}$ имеет вид $y = x^{(1)}x^{(2)}$. Тогда легко проверить, что $g_1 \equiv g_2 \equiv 1/4$, $U_1' = (1,1)$, $U_2' = (1,-1)$ точно восстанавливают функцию регрессии.

Этот же пример использован и для иллюстрации работы предлагаемого алгоритма при наличии выборки.

Другие возможные подходы. В отличие от работы [229], где делается попытка прямой аппроксимации функции регрессии, будем искать подпространство span $(U_1, \ldots, U_q)$, для которого достигает максимума значение ПИ:

$$Q_{\text{рег}}(U, X) = |\Sigma_Z|^{1/2} \int (\psi(y, Z) - \varphi(y) f(Z))^2 \, dZ dy, \quad (19.48)$$

где $\psi(y, Z)$ — плотность совместного распределения случайных величин $Z = (z^{(1)}, \ldots, z^{(q)})'$, $z^{(i)} = U_i' X$; $f(z)$ — маргинальная плотность распределения только $\mathbf{Z}$; $\varphi(y)$ — маргинальная плотность распределения y; Σ_Z — матрица ковариаций $\mathbf{Z}$.

ПИ (19.48) инвариантен относительно линейных преобразований Z, поэтому без ограничения общности можно считать, что компоненты вектора Z некоррелированы. В случае махаланобисовой метрики в пространстве $\Pi^p(X)$ это эквивалентно обычной попарной ортогональности векторов U_i, поэтому без ограничения общности можно считать, что $|\Sigma_Z| = 1$.

ПИ (19.48) является мерой расхождения модели «случайная величина y независима от Z» с ситуацией, имеющей место на самом деле. Максимизируя (19.48), ищут подпространство, где это расхождение максимально, т. е. такое, где y наиболее сильно зависит от Z.

19.8. Восстановление плотности и связь с томографией

19.8.1. Оценка плотности методом целенаправленного проецирования. Пусть имеется выборка $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ p-мерных наблюдений объема n. Опишем итерационную процедуру получения оценки плотности $f(X)$ в виде $f_0(X) \times \times \Pi g_m(\theta_m' X)$. Здесь $f_0(X)$ — начальная плотность, которая задается вместе с некоторой p-мерной выборкой $\mathbf{X}^{(n_0)} = (X_{10}, \ldots, X_{n_0 0})$ из нее, $n_0 \gg n$. На M-м шаге процедуры строится оценка в виде плотности

$$f_M(X) = f_{M-1}(X) g_M(\theta_M' X) \quad (19.49)$$

вместе с выборкой $\mathbf{X}^{n_M} = (X_{1M}, \ldots, X_{n_M M})$ из нее. Поправочная функция $g_M(y)$ и направление $\theta_M \in R^p$, $||\theta_M|| = 1$, вы-

бираются так, чтобы они минимизировали значение функционала относительной энтропии

$$H\left(f(X), \widetilde{f}(X)\right) = \int \log \frac{f(X)}{\widetilde{f}(X)} f(X)\, \mathrm{d}X$$

в классе всех плотностей $\widetilde{f}(X)$ вида $f_{M-1}(X)\, g(\theta' X)$. Имеем

$$H(f(X),\ f_{M-1}(X)\, g(\theta' X)) = H(f(X),\ f_{M-1}(X)) -$$

$$-\int \log g(\theta' X)\, f(X)\, \mathrm{d}X = H(f(X),\ f_{M-1}(X)) -$$

$$-\int \log g(y)\, \breve{f}(y,\ \theta)\, \mathrm{d}y, \qquad \text{где } \breve{f}(y,\theta) = \int\limits_{\theta' X = y} f(X)\, \mathrm{d}X.$$

Далее, из условия $\int \widetilde{f}\, \mathrm{d}X = 1$ следует, что $\int \breve{f}_{M-1}(y,\ \theta) \times \times g(y)\, \mathrm{d}y = 1$.

Положим $W(\theta, g(y)) = \int \log g(y)\, \breve{f}(y, \theta)\, \mathrm{d}y$.

Таким образом, $g_M(y)$ и θ_M являются решением задачи:

найти

$$\arg\max_{\theta,\ g} W(\theta,\ g(y)) \tag{19.50}$$

при условии $\int \breve{f}_{M-1}(y,\ \theta)\, g(y)\, \mathrm{d}y = 1$.

При каждом фиксированном θ задача (19.50) будет стандартной вариационной задачей, решением которой является функция

$$g_\theta(y) = \breve{f}(y,\ \theta)/\breve{f}_{M-1}(y,\ \theta). \tag{19.51}$$

Таким образом, $g_M(y) = g_{\theta_*}(y)$, где θ_* — решение следующей задачи целенаправленного проецирования:

найти

$$\arg\max_{\theta} W(\theta,\ \breve{f}(y,\ \theta)/\breve{f}_{M-1}(y,\ \theta)). \tag{19.52}$$

Наряду с выборкой $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$, по предположению индукции, имеется выборка $\mathbf{X}^{n_{M-1}} = (X_{1,\,M-1}, \ldots, X_{n_{M-1},\,M-1})$.

Опишем, как при помощи этих выборок получить оценку значения функционала $W(\theta, f(y, \theta))/f_{M-1}(y, \theta))$ каждого θ.

Фиксируем некоторый алгоритм оценки $\widetilde{\varphi}(y)$ плотности $\varphi(y)$ одномерной случайной величины по выборке $y_1, \ldots, y_N$. Например,

$$\widetilde{\varphi}(y) = \frac{1}{N} \sum^{N} k(y - y_i;\ h),$$

где $k(y; h) = \begin{cases} \frac{1}{h}, & \text{если } |y| \leqslant \frac{h}{2}; \\ 0, & \text{если } |y| > \frac{h}{2}. \end{cases}$

Тогда можно положить

$$\widetilde{g}_{\theta}(y) = \frac{n_{M-1} \sum^{n} k(y - \theta' \mathbf{X}_i;\ h)}{n \sum^{n} k(y - \theta' \mathbf{X}_{l,\, M-1};\ h)}. \qquad (19.53)$$

В (19.53) остается свободным параметр h. В [631] рекомендуется h выбирать зависящим от y так, чтобы условию $|y - \theta' X_{l,\, M-1}| \leqslant \frac{h}{2}$ удовлетворяло ровно $\alpha \cdot n_{M-1}$ точек $X_{l,\, M-1}$, где $\alpha = \text{const}$, например, 0,1; 0,05 и т.д. Тогда (19.53) примет вид:

$$\widetilde{g}_{\theta}(y) = \frac{1}{\alpha n} \sum^{n} k(y - \theta' X_l;\ h(y)). \qquad (19.54)$$

В качестве оценки функционала $W(\theta, g_{\theta}(y))$ можно взять функционал

$$\widetilde{W}(\theta) = \frac{1}{n} \sum^{n} \log \widetilde{g}_{\theta}(\theta,\ X_l) = \log \frac{1}{\alpha n} +$$

$$+ \sum_{i=1}^{n} \log \sum_{l=1}^{n} k(\theta'(\mathbf{X}_l - \mathbf{X}_i);\ h(\theta' \mathbf{X}_l)).$$

Решив теперь задачу:
найти

$$\arg\max_{\theta} \widetilde{W}(\theta), \qquad (19.55)$$

получаем оценку для θ_M и, следовательно, $g_M(y)$. Для завершения шага осталось построить выборку $\mathbf{X}^{n_M} = (X_{1M},$

..., $X_{n_M,\,M}$) из распределения $f_M(X)$. Воспользуемся следующим общим фактом [231], [258].

Пусть $f_0(X)$ и $f_1(X)$ — плотности распределения p-мерных случайных векторов; $\mathbf{X} = (X_1, \ldots, X_N)$ — выборка из распределения $f_0(X)$. Тогда если $r(X) = f_1(X)/f_0(X)$ — ограниченная функция, то следующий алгоритм просеивания позволяет получить из выборки $\mathbf{X}$ выборку $\widetilde{\mathbf{X}}$ из распределения $f_1(X)$.

Положим $\gamma = \max\limits_X r(X)$. Пусть $(u_1, \ldots, u_N)$ — выборка из равномерного распределения на интервале $[0, 1]$. Тогда наблюдение $X_i \in \mathbf{X}$ включается в $\widetilde{\mathbf{X}}$, если $u_i\gamma \leqslant r(X_i)$, и выбрасывается из рассмотрения в противном случае.

Из (19.49) и (19.51) теперь получаем:

$$r_M(X) = \frac{f_M(X)}{f_{M-1}(X)} = \frac{\widetilde{f}(\theta_M' X)}{\widetilde{f}_{M-1}(\theta_M' X)}.$$

Взяв согласно (19.54) в качестве оценки функции $r_m(X)$ функцию

$$\widetilde{g}_M(\theta' X) = \frac{1}{\alpha n} \sum k(\theta'(X - X_i);\ h(\theta' X)),$$

при помощи описанного выше алгоритма просеиваем выборку $\mathbf{X}^{n_{M-1}}$ из распределения $f_{M-1}(X)$ и получаем выборку $\mathbf{X}^{n_M}$ из распределения $f_M(X)$.

Таким образом, M-й шаг процедуры описан. В качестве начальной плотности $f_0(X)$, если нет дополнительной информации, обычно берется p-мерное нормальное распределение $N(a, \Sigma)$, где a и Σ — оценки по выборке $\mathbf{X}^{(n)}$ среднего значения и ковариационной матрицы.

19.8.2. Вычислительная томография и прикладная статистика. Термин «томография» (томо — сечение, слой; графия — описание) возник впервые в рентгенографии в начале нашего века и относился к восстановлению плотности сечения $f_\Pi(X)$ по ее проекциям аналоговым способом.

В общих чертах схема рентгеновской томографии следующая [162]. Пусть B — некоторое тело, например голова пациента, и $f(X)$ — плотность этого тела в точке $X \in R^3$. Если на B вдоль прямой ξ направить тонкий пучок рентгеновских лучей, то измеряемую величину $\log I_0/I$ можно с приемлемой точностью считать равной интегралу от $f(X)$ вдоль прямой ξ, где I_0 и I — интенсивности пучка до его попадания на тело B и после его выхода из B соответственно. Заставив источник рентгеновского излучения и детектор

двигаться так, чтобы соединяющая их прямая ξ находилась все время в плоскости Π, получаем возможность измерить проекцию $\check{f}_\Pi(y, \theta)$ плотности сечения $f_\Pi(X)$ тела B плоскостью Π, где θ — направление, ортогональное ξ в плоскости Π. Технические возможности рентгеновского томографа позволяют для фиксированной плоскости Π получить достаточно большой набор проекций $\check{f}_\Pi(y, \theta_1), \ldots, \check{f}(y, \theta_M)$. Ясно, что на самом деле измеряется не функция $\check{f}_\Pi(y, \theta_m)$, а набор ее значений $\check{f}_\Pi(y_k, \theta_m)$, $k = 1, \ldots, K$.

Основная задача томографии:

восстановить плотность сечения $f_\Pi(X)$, по ее проекциям $\check{f}_\Pi(y, \theta_1), \ldots, \check{f}_\Pi(y, \theta_M)$, точнее, восстановить $f_\Pi(X)$ по массиву данных $\{\check{f}_\Pi(y_k, \theta_m), k = 1, \ldots, K, n = 1, \ldots, N\}$. (19.56)

Аналоговый способ давал такую оценку $\tilde{f}_\Pi(X)$, которая не всегда позволяла с необходимой точностью решить основную задачу. Только соединение устройства для получения проекций $\check{f}_\Pi(y, \theta_m)$ с ЭВМ и создание соответствующего математического обеспечения позволило решить эту задачу с нужной для прикладных целей точностью. Рождение вычислительной рентгеновской томографии относится к началу 70-х годов. В настоящее время вычислительная томография — область научной и прикладной деятельности, в которой, с одной стороны, изучаются способы получения проекции $\check{f}_\Pi(y, \theta)$ на основе того или иного способа взаимодействия проникающего излучения (не обязательно рентгеновского) с телом, а с другой стороны, развиваются методы и программно-алгоритмическое обеспечение решения задачи (19.56).

Рассмотрим следующую статистическую задачу:

пусть имеется набор выборок $Y^{n_1} = (y_{11}, \ldots, y_{n,1}), \ldots, Y^{nM} = (y_{1,M}, \ldots, y_{nM})$ одномерных наблюдений. Восстановить плотность $f(X)$, $X \in R^p$, если известно, что Y^{n_m} — выборка из распределения $\check{f}(y, \theta_m)$ для каждого $m = 1, \ldots, M$, $\theta_m \in R^p$, $\|\theta_m\| = 1$, где $\check{f}(y, \theta_m) = \int\limits_{\theta_m X = y} f(X)\mathrm{d}X$. (19.57)

Задача (19.57) тесно связана с задачей томографии (19.56). Действительно, если дана выборка $\mathbf{X}^n = (X_1, \ldots, X_n)$ p-мерных наблюдений, то, положив $Y^{n_m} = (\theta_m' X_1, \ldots,$

$\theta'_m X_n$), оказываемся в условиях задачи (19.57) для любого набора направлений $\theta_1, \ldots, \theta_M$. С другой стороны, если даны выборки Y^m, $m = 1, \ldots, M$, то, восстановив по выборке Y^m плотность $\tilde{f}(y, \theta_m)$, оказываемся в условиях задачи (19.56).

19.8.3. Алгоритм восстановления плотности по ее проекциям на основе принципа минимальной вариабельности. Опишем теперь известный алгоритм из математического обеспечения томографии [162] как алгоритм восстановления многомерной плотности $f_*(X)$ по ее одномерным проекциям $\tilde{f}_*(y, \theta_1), \ldots, \tilde{f}_*(y, \theta_M)$.

В томографии, естественно, рассматриваются только плотности, сосредоточенные в ограниченных областях, поэтому будем считать, что $f_*(X)$ обращается в 0 вне шара $D \subset R^p$. Обозначим через $L_2(D)$ L_2 гильбертово пространство функций $f(X)$, $X \in R^p$, $\int\limits_D f^2(X)\,dX = \|f\|^2 < \infty$ со скалярным произведением $(f_1, f_2) = \int\limits_D f_1(X) f_2(X)\,dX$ и через $L_2(m) \subset L_2$ подпространство функций, таких, что

$$\tilde{f}(y, \theta_m) = \int\limits_D \hat{f}(X)\,\delta(\theta'_m X - y)\,dX = \tilde{f}_*(y, \theta_m). \qquad (19.58)$$

Так как $\tilde{f}_*(y, \theta_m)$ является плотностью распределения, то из (19.58) следует, что $\int\limits_D f(X)dX \quad 1$ для всех $f \in L_2(m)$.

Обозначим через π_m ортогональный проектор из L_2 в $L_2(m)$. Напомним, что, по определению,

$$\pi_m(f) = \arg\min_{f_1 \in L_2(m)} \|f - f_1\|^2.$$

Оператор π_m задается формулой

$$\pi_m(f)(X) = f(X) + [\tilde{f}_*(\theta'_m X, \theta_m) - \tilde{f}(\theta'_m X, \theta)] \times$$

$$\times \frac{R(X; \rho)}{\tilde{R}(\theta'_m X, \theta_m)}, \qquad (19.59)$$

где $R(X, \rho)$ — равномерное распределение в шаре D радиуса ρ.

Используя операторы π_m, можно для любого $f_0 \in L_2$ построить последовательность

$$f_0, f_1 = \pi_1 f_0, \ldots, f_m = \pi_m f_{m-1}, \ldots, \qquad (19.60)$$

где $\pi_{m_1} \equiv \pi_{m_2}$, если $m_1 - m_2$ делится на M, которая будет сходиться к решению следующей задачи:

найти

$$\widetilde{f}(x) = \arg\min_{f \in \bigcap_1^m L_2(m)} \|f - f_0\|^2. \qquad (19.61)$$

Решение задачи (19.61) и берется в качестве оценки плотности $f(X)$ с данными проекциями $\widetilde{f}(y, \theta_1), \ldots, \widetilde{f}(y, \theta_M)$.

Используя явную формулу (19.59) для проектора π_m, получаем из (19.60) описание алгоритма построения оценки $\widetilde{f}(X)$. В качестве начального приближения $f_0(X)$, если нет дополнительной информации, обычно берется равномерное распределение $R(X; \rho)$. В этом случае получается оценка $\widetilde{f}(X)$, которая среди всех $f \in \bigcap_1^m L_2(m)$ имеет наименьшую вариабельность, т. е. доставляет минимум функционалу $\int_D f^2 \, dX$ на $\bigcap_1^m L_2(m)$.

Из (19.59) и (19.60) следует

$$\|f_{m-1} - \pi_m f_{m-1}\|^2 = \int_D \left(\widetilde{f}_*(\theta_m' X, \theta_m) - \widetilde{f}_{m-1}(\theta_m' X, \theta_m)\right)^2 \times$$

$$\times \frac{R(X, \rho)^2}{R(\theta_m' X, \theta_m)^2} \, dX = c \int_{-\rho}^{\rho} \frac{(\widetilde{f}_*(y, \theta_m) - \widetilde{f}_{m-1}(y, \theta_m))^2}{\widetilde{R}(y, \theta_m)} \, dy,$$

где $c = \text{const}$.

Таким образом, можно использовать аргументы целенаправленного проецирования для получения более быстрой модификации описанного алгоритма.

Положим

$$F(\theta_m, f) = \int_{-\rho}^{\rho} \frac{(\widetilde{f}_*(y, \theta_m) - \widetilde{f}(y, \theta_m))^2}{\widetilde{R}(y, \theta_m)} \, dy. \qquad (19.62)$$

Пусть уже построены приближения $f_0, f_1, \ldots, f_k$. Взяв функционал $F(\theta_m, f_k)$ в качестве критерия выразительности проекции $\widetilde{f}_*(y, \theta_m)$ относительно приближения f_k, найдем

$$m(k) = \arg\max_{1 \leqslant m \leqslant M} F(\theta_m, f_k)$$

и зададим следующее приближение формулой $f_{k+1}(X) = \pi_{m(k)} f_k(X)$.

Замечание. Функционал (19.62) можно использовать в разведочном анализе для нахождения выразительных проекций данной выборки $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ p-мерных наблюдений, считая k-й по важности выразительной проек-

цией проекцию $(\theta_k' X_1, \ldots, \theta_k' X_n)$, для которой оценка функционала $F(\theta_k, \hat{f}_{k-1})$, $k \geqslant 1$ достигает своего максимального значения. Таким образом, решая задачу восстановления плотности по выборке указанным выше алгоритмом, по ходу получения оценок плотности $\hat{f}_{k-1}$ будем получать и соответствующие выразительные проекции

19.8.4. Алгоритм восстановления плотности по ее проекциям на основе принципа максимума энтропии. Пусть $\check{f}_*(y, \theta_1)$, $\ldots, \check{f}_*(y, \theta_M)$ — данный набор одномерных проекций искомой многомерной плотности $f_*(X)$

Обозначим через L_1^+ пространство плотностей $f(X)$ в D и через $L_1^+(m) \subset L_1^+$ — подпространство плотностей $f(X)$, таких, что

$$\check{f}(y, \theta_m) = \int_D f(X)\, \delta(\theta_m' X - y)\, dX = \check{f}_*(y, \theta_m),$$

$$\int_D f(X)\, dX = 1.$$

Здесь, в отличие от п. 19.8.3, можно считать D неограниченным подмножеством в R^p, в частности, D может совпадать с R^p.

В основе конструкции алгоритма из п.19.8.3 лежит геометрия гильбертова пространства L_2. Используя геометрию, задаваемую в L_1^+ функционалом относительной энтропии

$$H(f, f_1) = \int_D \log \frac{f}{f_1} f\, dX,$$

несимметричную пифагорову геометрию информационного уклонения, в терминологии Н. Н. Ченцова [165], можно тем же способом построить проекционный алгоритм восстановления плотности.

Определим проектор τ_m из L_1^+ в $L_1^+(m)$ как оператор, ставящий в соответствие плотности $f \in L_1^+$ плотность

$$\tau_m(f) = \arg \min_{f_1 \in L_1^+(m)} H(f_1, f).$$

Оператор τ_m задается формулой

$$\tau_m(f) = f(X)\, \frac{\check{f}_*(\theta_m' X, \theta_m)}{\check{f}(\theta_m' X, \theta_m)}.$$

которая вытекает из соотношения

$$H(f_1(X),\ f(X)) = H\left(f_1(X),\ f(X)\ \frac{\check{f}_*(\theta'_m X,\ \theta_m)}{\check{f}(\theta'_m X,\ \theta_m)}\right) +$$

$$+\ H(f_1(y,\ \theta_m),\ \check{f}(y,\ \theta_m)),$$

верного для всех $f_1 \in L_1^+(m)$ и $f \in L_1^+$.
При помощи операторов τ_m так же, как и в п. 19.8.3, для любой начальной плотности $f_0 \in L_1^+$ строится последовательность

$$f_0,\ f_1 = \tau_1 f_0,\ \ldots,\ f_m = \tau_m f_{m-1},\ \ldots \tag{19.63}$$

Цель алгоритма (19.63) — дать в качестве оценки плотности $f_*(X)$ решение задачи:
найти

$$\tilde{f}(x) = \arg \min_{f \in \cap L_1^+(m)} H(f,\ f_0). \tag{19.64}$$

Пусть $f_*(X)$ сосредоточено в шаре D радиуса ρ. Тогда, взяв в качестве $f_0(X)$ равномерное распределение $R(X;\ \rho)$, получаем

$$H(f,\ R(X,\ \rho)) = \int_D \log \frac{f}{R} f \mathrm{d}X = \int_D \log f \cdot f \mathrm{d}X + \mathrm{const} =$$

$$= \mathrm{const} - H(f),$$

т. е. в этом случае задача (19.64) сводится к задаче:
найти

$$\tilde{f}(X) = \arg \max_{f \in \cap L_1^+(m)} H(f). \tag{19.65}$$

Пусть ранг системы векторов $\theta_1, \ldots, \theta_m$, $m = 1, \ldots, M$, $\theta_m \in R^p$ не меньше p. Без ограничения общности в этом случае можно считать, что матрица Θ, составленная из векторстолбцов $\theta_1, \ldots, \theta_p$, является невырожденной. Тогда, взяв в качестве начального приближения $f_0(X) = \prod_1^p \check{f}_*(\theta'_i X,\ \theta_i) |\det \Theta|$, получаем:

$$H(f,\ f_0) = -H(f) - \log |\det \Theta| + \sum_{l=1}^{p} H(\check{f}_*(y,\ \theta_l)),$$

т. е. $H(f,\ f_0) = -H(f) + \mathrm{const}$ для любой плотности $f \in L_1^+(m)$. Следовательно, и в этом случае задача (19.64)

сводится к задаче (19.65), но теперь уже без дополнительного предположения о том, что $f_*(X)$ сосредоточено в шаре D. Из (19.63) следует

$$H(f_m, f_{m-1}) \leqslant H(\widetilde{f}_*(y, \theta_m), \widetilde{f}_{m-1}(y, \theta_m)).$$

Таким образом, как и в п. 19.8.3, можно использовать аргументы целенаправленного проецирования для модификации алгоритма.

Положим $\Phi(\theta_m, f) = H(\widetilde{f}_*(y, \theta_m), \widetilde{f}(y, \theta_m))$. (19.66) Пусть уже построены приближения $f_0, f_1, \ldots, f_k$. Взяв функционал $\Phi(\theta_m, f_k)$ в качестве критерия выразительности проекции $\widetilde{f}_*(y, \theta_m)$ относительно приближения f_k, найдем

$$m(k) = \arg\max_{1 \leqslant m \leqslant M} \Phi(\theta_m, f_k)$$

и зададим следующее приближение формулой $f_{k+1}(X) = \tau_{m(k)} f_k(X)$.

Замечание. Если использовать функционал (19.66) в разведочном анализе для нахождения наиболее выразительных проекций данной выборки $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ среди всех проекций $(\theta_m' X_1, \ldots, \theta_m' X_n)$, где θ_m пробегает фиксированный список направлений, а именно так и бывает при численной реализации алгоритмов ЦП, то видно, что в этом случае алгоритм оценки плотности, данный в п. 19.8.1, совпадает с только что рассмотренным модифицированным алгоритмом.

19.9. Некоторые вопросы вычислительной реализации и практические приемы целенаправленного проецирования

19.9.1. Вычислительные процедуры. Для части ПИ вычислительные процедуры рассмотрены в соответствующих параграфах (см. § 19.5, 19.7, 19.8). Здесь же остановимся на ПИ типа, рассмотренного в § 19.4. Для реализации вычислительной процедуры, когда задана выборка $\mathbf{X}^{(n)}$, необходимо уметь вычислять оценку ПИ (см. § 19.4) для любой проекции по выборке и градиент или матрицу вторых производных от этой оценки.

Оценка значения ПИ. Возможно несколько способов оценки функционалов вида, рассмотренного в § 19.4, от плотности проекций $z = U'X$. Во-первых, можно несколькими способами непараметрически оценить саму плотность (ядерная оценка, оценка по методу k-ближайших соседей,

гистограммная оценка и т.д.) и затем оценить сам функционал [164]. Другой метод основан на использовании так называемых gaps-статистик [208, 326]. Этот подход и будет далее рассмотрен. Пусть $z_i = (U'X_i)$ $(i = \overline{1, n})$ — проекции векторов из выборки на вектор U, а $z_{(1)}, \ldots, z_{(n)}$ — соответствующие порядковые статистики (вариационный ряд; см., например, [11]). Образуем gaps-статистики вида

$$\Delta_{i,r} = z_{(i+r)^-} - z_{(i-r)^+}, \tag{19.67}$$

где $(i + r)^- = \min(n, i + r)$; $(i - r)^+ = \max(1, i - r)$, r — целое число $(r < n/2)$.

Можно показать, что сумма

$$E_{\beta, r} = \left(\frac{2r}{n}\right)^{\beta} \Sigma \Delta_i^{-\beta}/n \tag{19.68}$$

является оценкой для $\mathbf{E}_z f^{\beta}(z)$. Оценка (19.68) асимптотически нормальна и состоятельна при некоторых условиях на скорость роста r с ростом объема выборки n.

Величина окна r играет роль, аналогичную роли параметра сглаживания для ядерных оценок или числа соседей для оценки по методу k-ближайших соседей. Как уже указывалось, она должна возрастать с ростом n. Некоторые соображения о выборе значения r на практике приведены ниже. Окончательной оценкой ПИ (19.4) будет

$$Q(U, \mathbf{X}^{(n)}) = \widehat{s^{\beta}} E_{\beta, r}. \tag{19.69}$$

Дальше, поскольку ПИ (19.69) афинноинвариантен, будем считать, что предварительно перешли к махаланобисовой метрике. Это дает следующее преимущество — условие S-ортогональности в лемме 19.1 заменяется обычной ортогональностью и, кроме того, облегчает аналитическое вычисление направления градиента для (19.69).

Вычисление градиента. Градиент ПИ (19.69) получается прямым дифференцированием $Q(U, \mathbf{X}^{(n)})$ по U. При этом нужно учесть, что направление градиента должно быть ортогонально вектору U. Так как производная от s^{β} по U дает только составляющую, параллельную U, то направление градиента будет совпадать с направлением ортогональной к U составляющей $\partial E_{\beta, r}/\partial U$:

$$\operatorname{grad} Q|(U, \mathbf{X}^{(n)}) \sim ORT_U (\partial E_{\beta, r}/\partial U).$$

Выражение же для $(\partial E_{\beta,\, r}/\partial U)$:

$$\partial E_{\beta,\, r}/\partial U = \left(\frac{r}{\beta}\right)^{\beta} (-\beta) \sum_{i=1}^{n} \Delta_{i,\, r}^{-\beta-1} (X_{(i+r)} - X_{(i-r)}), \tag{19.70}$$

где $X_{(j)}$ — вектор из выборки $\mathbf{X}^{(n)}$, проекция которого дает j-ю порядковую статистику, т. е. $z_{(j)} = U' X_{(j)}$

Зная направление градиента, можно теперь строить различные оптимизационные процедуры.

19.9.2. Практические рекомендации при проведении ЦП. **Выбор величины окна** r. При программной реализации управление значением этого параметра должно быть в той или иной степени доступно пользователю. Оптимальное значение параметра r зависит от объема выборки n, параметра β и неизвестной функции плотности распределения компонентов смеси. В реальной ситуации, когда модель (19.2) может выполняться лишь приближенно, теоретический выбор еще более затруднен. Имеется лишь некоторое предварительное впечатление для величины r, полученное на основе статистического моделирования с использованием смесей нормальных распределений. Так, при $n = 100$ диапазон «удачных» значений r будет 5—15, при $n = 200 - 10-30$. Впрочем, влияние величины r не слишком значительно. Все же рекомендуется провести вычисления с разными значениями r. Это позволяет увеличить и вероятность попадания в глобальный максимум функции (19.69).

Переход к махаланобисовой метрике. Как указано в п. 19.9.1, целесообразно перейти перед проведением ЦП к махаланобисовой метрике, так чтобы общая ковариационная матрица выборки стала единичной ($\mathbf{S} = \mathbf{I}_p$). Это позволяет использовать обычное условие ортогональности вместо S-ортогональности. В программе, реализующей ЦП, при использовании ПИ вида (19.4) такой переход должен делаться принудительно, без участия пользователя.

Сокращение размерности перед использованием процедур ЦП. Процедуры ЦП целесообразно сочетать с предварительным сокращением размерности по методу главных компонент. Необходимо удалить компоненты с малой дисперсией — подпространство, где отсутствует разброс точек, не может содержать какой-либо структуры. Контроль за количеством отбрасываемых компонент может осуществляться как пользователем, так и самой программой. Как и при выборе параметра сглаживания, имеет смысл провести несколько отсчетов с разным количеством отброшенных главных компонент.

Подавление влияния аномальных наблюдений. Эти наблюдения сильно влияют на результаты ЦП практически при использовании любых ПИ. Так, при наличии аномальных наблюдений проекции, получаемые с использованием ПИ (19.4), в основном будут выделять эти аномальные наблюдения, но не кластеры. Поэтому целесообразно сначала провести ЦП для выделения аномальных наблюдений с помощью простой процедуры из § 19.5. Там же будут получены веса w_i для каждого из наблюдений X_i (см. пример 19.3). Дальше можно либо отбросить долю α наблюдений с минимальным весом (эта доля может иметь стандартное значение $\alpha = 0{,}05$ либо задаваться пользователем), либо перейти к взвешенной оценке ПИ. Например, для ПИ (19.4) можно заменить оценку (19.68) на

$$\tilde{E}_{\beta,\,r} = \left(\frac{r}{n}\right)^{\beta} \Sigma w_i \, \Delta_{i,\,r}^{\beta}. \qquad (19.71)$$

И использовать устойчивую оценку дисперсии s^2.

Соответственно меняется и градиент.

Сглаживание. В реальной практике распределения часто либо дискретны, либо содержат дискретную составляющую. Чтобы избежать вычислительных трудностей, связанных с тем, что величина $\Delta_{i,r}$ (19.67) обращается в нуль, можно использовать сглаженную величину $\tilde{\Delta}_{i,r} = \Delta_{i,r} + \delta$, где δ есть, например, $\delta = \gamma\,(z_{(n-1)} - z_{(2)})/n$, а γ — малая величина порядка 0,01.

ВЫВОДЫ

1. Техника ЦП основана на поиске небольшого числа q выразительных (информативных, интересных) линейных проекций исходных p-мерных данных ($p \gg q$) из условия максимизации некоторых функционалов (проекционных индексов). ПИ подбираются таким образом, чтобы в спроецированных данных сохранялась вся информация о структуре исходных многомерных данных.
2. Полученные проекции могут быть использованы либо для визуального анализа структур (если $q > 3$), либо производится агрегирование содержащейся в них информации для восстановления поверхности регрессии (см. §19.7) плотности распределения (см. § 19.8).
3. ПИ для поиска выразительных проекций конструируются на одном из следующих принципов: как мера отклонения от нормального распределения, как мера отклонения от ги-

потезы независимости (см. § 19.6, 19.7), как ПИ, максимизация которых порождает базис в дискриминантном подпространстве (см. § 19.2).

4. При практическом применении ЦП нужно, во-первых, по возможности сокращать размерность пространства переменных (например, используя метод главных компонент), во-вторых, подавлять влияние аномальных наблюдений.

Глава 20. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЦЕЛЕНАПРАВЛЕННОГО ПРОЕЦИРОВАНИЯ И ТОМОГРАФИЧЕСКИХ МЕТОДОВ АНАЛИЗА ДАННЫХ

20.1. Проекции многомерных распределений и их свойства

20.1.1. Основные определения. Рассмотрим евклидово пространство R^p размерности p. Проекцией из R^p в R^q, $q \leqslant p$ будем называть *линейное* отображение $\mathbf{A}$ из R^p на *все* R^q. Фиксировав в R^p и R^q ортонормированные базисы $\theta_1, \ldots, \theta_p$ и $\theta_1, \ldots, \theta_q$, можно задать проекцию $(p \times q)$-матрицей $\mathbf{A}$ ранга q, т. е. $\mathbf{AA}'$ будет невырожденной матрицей. Здесь $\mathbf{A}'$ — транспонированная матрица $\mathbf{A}$. Проекция называется *ортогональной*, если $\mathbf{AA}' = \mathbf{I}_q$ — единичная матрица. Важным частным случаем являются *одномерные* проекции, т. е. проекции из R^p в R^1. Они задаются формулой

$$\mathbf{A}X = \theta' X,$$

где $\theta' X = \sum_{i=1}^{p} \theta^{(i)} x^{(i)}$ — скалярное умножение в R^p.

Ортогональные одномерные проекции задаются векторами θ с $||\theta|| = (\theta'\theta)^{1/2} = 1$.

Проекцией распределения векторной величины в R^p, соответствующей проекции $\mathbf{A}$ из R^p в R^q называется *распределение* q-мерной величины, *индуцированное* проекцией $\mathbf{A}$. Например, если ξ — случайный вектор в R^p с плотностью распределения $f_\xi(x)$, то его проекция $\mathbf{A}\xi = \eta$ — случайный вектор в R^q с плотностью

$$f_{\mathbf{A}\xi}(Y) = \int_{\mathbf{A}X = Y} f_\xi(X)\, dX, \quad Y \in R^q, \quad X \in R^p. \tag{20.1}$$

20.1.2. Общие свойства проекции распределения. Пусть $\mathbf{Q}_p : R^p \to R^p$, $\mathbf{Q}_q : R^q \to R^q$ — невырожденные линейные отображения и $X_0 \in R^p$. Тогда

$$f_{\mathbf{Q}_p \xi + X_0}(X) = f_\xi(\mathbf{Q}_p^{-1}(X - X_0)) \,|\det \mathbf{Q}_p|^{-1}; \tag{20.2}$$

$$f_{\mathbf{A}(\mathbf{Q}_p \xi + X_0)}(Y) = \int\limits_{\mathbf{A}X = Y} f_\xi(\mathbf{Q}_p^{-1}(X - X_0)) |\det \mathbf{Q}_p|^{-1} \, dX =$$

$$= f_{\mathbf{A}\mathbf{Q}_p \xi}(Y - \mathbf{A}X_0); \tag{20.3}$$

$$f_{\mathbf{Q}_q \mathbf{A}\xi}(Y) = f_{\mathbf{A}\xi}(\mathbf{Q}_q^{-1} Y) \,|\det \mathbf{Q}_q|^{-1}. \tag{20.4}$$

Для данной проекции $\mathbf{A} : R^p \to R^q$ рассмотрим симметрическую положительно определенную $(q \times q)$-матрицу $\mathbf{AA}'$. Пусть $\mathbf{C}$ — ортогональная матрица, составленная из собственных вектор-столбцов матрицы $\mathbf{AA}'$ и $\mathbf{\Lambda}$ — диагональная матрица $\|\lambda_1, \ldots, \lambda_q\|$, где $\lambda_i > 0$ — соответствующие собственные числа, т. е. $\mathbf{C}'(\mathbf{AA}')\,\mathbf{C} = \mathbf{\Lambda}$. Положим $\mathbf{B} = \mathbf{C}\mathbf{\Lambda}^{-1/2}\mathbf{C}$, где $\mathbf{\Lambda}^{-1/2} = \|\lambda_1^{-1/2}, \ldots, \lambda_q^{-1/2}\|$. Тогда

$$(\mathbf{BA})(\mathbf{BA})' = \mathbf{B}(\mathbf{AA}')\mathbf{B} = \mathbf{I}_q,$$

т. е. $\mathbf{BA}$ — матрица ортогональной проекции из R^p в R^q. Используя формулу (20.4), получаем, что проекция $f_{\mathbf{A}\xi}(Y)$ выражается через ортогональную проекцию $f_{\mathbf{BA}\xi}(Y)$. Формулу (20.1) в случае одномерных проекций $\mathbf{A}\xi = \theta'\xi = \sum\limits_{i=1}^{p} \theta^i \xi^i$ можно записать в виде преобразования Радона плотности $f_\xi(X)$ [163]:

$$f_\xi(y;\ \theta) = \int\limits_{\theta' X = y} f(X)\, dX = \int\limits_{R^p} f_\xi(X)\, \delta(\theta' X - y)\, dX, \tag{20.5}$$

где $\delta(y_0 - y)$ (δ-функция Дирака) — одномерная плотность, сосредоточенная в точке y_0. Формулы (20.2) и (20.3) перепишутся теперь в виде:

$$\tilde{f}_{\mathbf{Q}_p \xi + X_0}(y;\ \mathbf{a}) = \tilde{f}_\xi(y - \mathbf{a}' X_0)\, \mathbf{Q}_p' \mathbf{a}; \tag{20.6}$$

$$\tilde{f}_\xi(y,\ \lambda \mathbf{a}) = \frac{1}{|\lambda|}\, \tilde{f}_\xi\left(\frac{y}{\lambda},\ \mathbf{a}\right), \tag{20.7}$$

где λ — ненулевое число.

Рассмотрим характеристическую функцию $\varphi(t;\ \mathbf{a})$ случайной величины $y = \mathbf{a}'\xi$ [111]:

$$\varphi(t;\ \mathbf{a}) = \mathbf{E}(e^{ity}) = \int\limits_{-\infty}^{\infty} e^{ity}\, \tilde{f}_\xi(y,\ \mathbf{a})\, dy.$$

Имеет место формула

$$\varphi(t;\ \mathbf{a}) = \int\limits_{-\infty}^{\infty} e^{itg} \left(\int\limits_{R^p} f_\xi(X)\, \delta(\mathbf{a}' X - y)\, dX \right) =$$

$$= \int\limits_{R^p} f_\xi(X)\, e^{it(\mathbf{a}' X)}\, dX. \qquad (20.8)$$

Следовательно, $\varphi(1;\ \mathbf{a})$ как функция вектора $\mathbf{a} \in R^p$ является характеристической функцией p-мерного случайного вектора ξ. Так как $\varphi(1,\ \mathbf{a})$ рассчитывается по $\breve{f}_\xi(y,\ \mathbf{a})$, то из теоремы обращения характеристической функции [129] получаем: распределение p-мерного вектора полностью определяется распределениями его одномерных проекций.

Этот важнейший результат в теории преобразования Радона называется *теоремой о связи преобразований Радона и Фурье*, *теоремой о проекциях и сечениях* [162, 163], а в многомерном статистическом анализе — *теоремой Крамера и Волда* [129]. В теории преобразования Радона получены явные формулы, выражающие $f_\xi(X)$ через семейство $\breve{f}_\xi(y,\ \mathbf{a})$, где $\mathbf{a}$ пробегает множество $S^{p-1} = \{\mathbf{a} \in R^p,\ \|\mathbf{a}\| = 1\}$, а также $f_\xi(X)$ через семейство $f_{\mathbf{A}\xi}(y)$, где $\mathbf{A}$ пробегает множество ортогональных проекций из R^p в R^q.

Формула (20.8) описывает частный случай следующего общего свойства проекций $\breve{f}_\xi(y,\ \mathbf{a})$ плотности $f_\xi(X)$:

$$\int\limits_{-\infty}^{\infty} \varphi(y)\, \breve{f}_\xi(y,\ \mathbf{a})\, dy = \int\limits_{Rp} \varphi(\mathbf{a}' \xi)\, f_\xi(X)\, dX,$$

т. е.

$$\mathbf{E}(\varphi(y);\ \breve{f}_\xi(y,\ \mathbf{a})) = \mathbf{E}(\varphi(\mathbf{a}' X);\ f_\xi(X)). \qquad (20.9)$$

20.1.3. Свойства проекций дифференцируемых распределений. В тех случаях, когда плотность $f_\xi(X)$ дифференцируема, то ее градиент $\nabla_X f_\xi(X) = \left(\frac{\partial}{\partial x_1} f_\xi(X),\ \ldots,\ \frac{\partial}{\partial x_p} f_\xi(X) \right)$ выражается в терминах проекций $\breve{f}(y,\ \mathbf{a})$ формулой

$$(\mathbf{b}' \nabla_X f_\xi(X))\breve{}\,(y,\ \mathbf{a}) = (\mathbf{b}' \mathbf{a}) \frac{\partial}{\partial y} \breve{f}_\xi(y,\ \mathbf{a}), \qquad (20.10)$$

где $\mathbf{b}$ и $\mathbf{a}$ — любые ненулевые векторы из R^p. В частности, когда $\mathbf{b} = \mathbf{a}$ и $||\mathbf{a}|| = 1$, то

$$(\mathbf{a}' \nabla_x f_\xi(x))^{\sim}(y, \mathbf{a}) = \frac{\partial}{\partial y} \tilde{f}_\xi(y, \mathbf{a}). \tag{20.11}$$

Для описания связи между проекциями $\tilde{f}_\xi(y, \theta_1)$ и $\tilde{f}_\xi(y, \theta_2)$ для близких направлений θ_1 и θ_2 важна следующая формула:

$$\mathbf{b}' \nabla_\theta \tilde{f}_\xi(y, \theta) = -\frac{\partial}{\partial y}\{(\mathbf{b}' x) f_\xi(x)\}^{\sim}(y, \theta). \tag{20.12}$$

Для случайного вектора ξ в R^p с плотностью $f_\xi(x)$ обозначим через $\bar{X}_\xi(y, \mathbf{a})$ вектор в R^p, равный среднему среди векторов, лежащих на гиперплоскости $\mathbf{a}' x = y$, т. е.

$$\bar{X}_\xi(y, \mathbf{a}) = \frac{1}{\tilde{f}_\xi(y, \mathbf{a})} \int_{\mathbf{a}' x = y} x f_\xi(x)\, dx. \tag{20.13}$$

Тогда из (20.12) и (20.13) получаем:

$$\nabla_\mathbf{a} \tilde{f}_\xi(y, \mathbf{a}) = -\frac{\partial}{\partial y}(\bar{X}_\xi(y, \mathbf{a})\, \tilde{f}_\xi(y, \mathbf{a})). \tag{20.14}$$

Рассматривая теперь вектор $\mathbf{a}$ как p-мерный параметр распределения $\tilde{f}_\xi(y, \mathbf{a})$, составим для каждого $\mathbf{a}$ информационную матрицу Фишера $\mathbf{I}(\mathbf{a}; \xi)$ [11, с. 256]:

$$\mathbf{I}(\mathbf{a}; \xi) = \{I_{kl}(\mathbf{a}; \xi)\} = \mathbf{E}\,|\nabla_\mathbf{a} \log \tilde{f}_\xi(y, \mathbf{a}) \cdot \nabla_\mathbf{a}' \log \tilde{f}_\xi(y, \mathbf{a});$$
$$\tilde{f}_\xi(y, \mathbf{a})|.$$

Применяя (20.14), получаем:

$$\mathbf{I}(\mathbf{a}; \xi) = \mathbf{E}\left[\frac{\partial \bar{X}_\xi(y, \mathbf{a})\, \tilde{f}_\xi(y, \mathbf{a})}{\partial F_\xi(y, \mathbf{a})} \cdot \frac{\partial \bar{X}_\xi'(y, \mathbf{a})\, \tilde{f}_\xi(y, \mathbf{a})}{\partial F_\xi(y, \mathbf{a})}; \right.$$
$$\left. \tilde{f}_\xi(y, \mathbf{a})\right], \tag{20.15}$$

где $F_\xi(y, \mathbf{a})$ — функция распределения случайной величины $\eta = \mathbf{a}'\xi$.

Когда вектор $\mathbf{a}$ пробегает сферу $S^{p-1} = \{\mathbf{a} \in R^p, ||\mathbf{a}|| = 1\}$, получаем поле неотрицательно определенных симметрических матриц $\mathbf{I}(\mathbf{a}; \xi)$ на S^{p-1}. Это поле можно исполь-

зовать для построения критерия относительной выразительности направлений проецирования $\mathbf{a} \in S^{p-1}$. Положим

$$\Phi(\mathbf{a}) = \int_{S_{\mathbf{a}}^{p-2}} (\mathbf{b}' I(\mathbf{a};\ \xi)\, \mathbf{b})\, d\mathbf{b}, \tag{20.16}$$

где $S_{\mathbf{a}}^{p-2} = \{\mathbf{b} \in R^p;\ \|\mathbf{b}\| \quad 1,\ \mathbf{b}'\mathbf{a} = 0\}$. Содержательно $\Phi(\mathbf{a})$ указывает, какова усредненная по $\mathbf{b}$ чувствительность распределения $\widetilde{f}_{\xi}(y,\ \mathbf{a})$ к изменениям направления проецирования вида $\mathbf{a} + \varepsilon\mathbf{b}$ для малых ε.

Из (20.15) получаем:

$$\Phi(\theta) = \mathrm{E}\left(\int_{S_{\theta}^{p-2}} \left(\mathbf{b}\, \frac{\partial \overline{X}_{\xi}(y,\ \mathbf{a})\, f_{\xi}(y,\ \mathbf{a})}{\partial F_{\xi}(y,\ \mathbf{a})} \right)^2 d\mathbf{b};\ \widetilde{f}_{\xi}(y,\ \theta) \right).$$

Используя теперь, что если $\|\theta\| = 1$, то $\overline{X}_{\xi}(y,\ \theta) = y\theta + \overline{X}_{\xi 1}(y,\ \theta)$, где $\theta'\overline{X}_{\xi 1}(y,\ \theta) = 0$ и формулу

$$\int_{S_{\mathbf{a}}^{p-2}} (\mathbf{b}' Z)^2\, d\mathbf{b} = \|Z\|^2 - (Z'\mathbf{a})^2,$$

верную для всех $Z \in R^p$, получаем:

$$\Phi(\mathbf{a}) = \mathrm{E}\left(\left\| \frac{\partial \overline{X}_{\xi}'(y,\ \mathbf{a})\, \widetilde{f}_{\xi}(y,\ \mathbf{a})}{\partial F_{\xi}(y,\ \mathbf{a})} \right\|^2;\ \widetilde{f}_{\xi}(y,\ \mathbf{a}) \right) \tag{20.17}$$

Пример 20.1. Пусть ξ — нормальный p-мерный вектор $N(X_0, \Sigma)$.

Тогда согласно (20.17) получаем:

$$\Phi(\mathbf{a}) = \mathrm{E}\left(\left\| \left(\frac{\Sigma\mathbf{a}}{\sigma_{\mathbf{a}}} - \sigma_{\mathbf{a}}\mathbf{a} \right) \left(1 - \left(\frac{y - y_{\mathbf{a},0}}{\sigma_{\mathbf{a}}} \right)^2 \right) - \right.\right.$$

$$\left.\left. - \left(\frac{X_0 - y_{\mathbf{a}\ 0}\,\mathbf{a}}{\sigma_{\mathbf{a}}} \right) \left(\frac{y - y_{\mathbf{a},0}}{\sigma_{\mathbf{a}}} \right) \right\|^2;\ N(y_{\mathbf{a},0},\ \sigma_{\mathbf{a}}^2) \right)$$

$$= 2 \left\| \frac{\Sigma\mathbf{a}}{\sigma_{\mathbf{a}}} - \sigma_{\mathbf{a}}\mathbf{a} \right\|^2 + \left\| \frac{X_0 - \mathbf{a}'X_0\,\mathbf{a}}{\sigma_{\mathbf{a}}} \right\|^2$$

В частности, если

1) $X_0 \neq 0$ и $\Sigma = I_p$ — единичная матрица, то

$$\Phi(\mathbf{a}) = \| X_0 - (\mathbf{a}' X_0)\,\mathbf{a} \|^2 = \|X_0\|^2 - (\mathbf{a}'X_0)^2,$$

т. е. критерий $\Phi(\mathbf{a})$ принимает минимальное значение, если $\mathbf{a} = \frac{X_0}{\|X_0\|}$, и максимальное значение, если $\mathbf{a}' X_0 = 0$;
2) $X_0 = 0$, тогда

$$\Phi(\mathbf{a}) = 2\left\| \frac{\Sigma \mathbf{a}}{\sigma_{\mathbf{a}}} - \sigma_{\mathbf{a}} \mathbf{a} \right\|^2 = 2\left(\frac{\| \Sigma \mathbf{a} \|^2}{\sigma_{\mathbf{a}}^2} - \sigma_{\mathbf{a}}^2 \right),$$

т. е. $\Phi(\mathbf{a}) \quad 0$ тогда и только тогда, когда $\Sigma \mathbf{a} = \sigma_{\mathbf{a}}^2 \mathbf{a}$, т.е. когда вектор проецирования совпадает с главной компонентой.

20.1.4. Связь многомерного распределения с его одномерной проекцией. Рассмотрим теперь насколько характеризует данное p-мерное распределение с плотностью $f_\xi(X)$ его единственная одномерная проекция $\tilde{f}(y, \mathbf{a})$.

Положим $Y(f_\xi(y, \mathbf{a})) = \{y \in R^1; \tilde{f}_\xi(y, \mathbf{a}) > 0\}$, $Y(f_\xi; \mathbf{a})$ называется носителем плотности $\tilde{f}_\xi(y, \mathbf{a})$.

Пусть $f_{\xi_1}(X)$ — некоторая плотность, удовлетворяющая относительно $f_\xi(X)$ и фиксированного $\mathbf{a}_0$, $\|\mathbf{a}_0\| = 1$ только условию

$$Y(f_\xi; \mathbf{a}_0) \subseteq Y(f_{\xi_1}; \mathbf{a}_0).$$

Тогда согласно свойству (20.9) функция

$$f(X) = \frac{f_{\xi_1}(X)}{\tilde{f}_{\xi_1}(\mathbf{a}_0' X, \mathbf{a}_0)} \tilde{f}_\xi(\mathbf{a}_0' X, \mathbf{a}_0)$$

задает плотность распределения, причем $\tilde{f}(y, \mathbf{a}_0) \equiv \tilde{f}_\xi(y, \mathbf{a}_0)$. Таким образом, единственная проекция $\tilde{f}_\xi(y, \mathbf{a}_0)$ определяет распределение $f_\xi(X)$ только с точностью до множителя $f_{\xi_1}(X)/\tilde{f}_{\xi_1}(\mathbf{a}_0' X, \mathbf{a}_0)$, где $f_{\xi_1}(X)$ — фактически произвольная плотность. Столь же малую информацию о распределении *общего вида* несет и любой конечный набор его проекций $\tilde{f}_\xi(y, \mathbf{a}_l)$, $l = 1, \ldots, L$

В связи с этим, как уже отмечалось выше (см. гл. 19), в задачах анализа многомерного распределения по его проекциям первостепенное значение имеет выбор *модели* этого распределения, либо *критерия*, при помощи которого среди всех распределений, имеющих данные проекции $\tilde{f}_\xi(y, \mathbf{a}_l)$, $l = 1, \ldots, L$, отбирается распределение, экстремальное по этому критерию. Алгоритмы решения таких задач рассмотрены в § 19.8.

20.2. Радиальные распределения

20.2.1. Основные понятия. Общие свойства радиальных распределений и их проекций. Рассмотрим класс многомерных распределений, смеси которых дают запас модельных законов распределения, достаточный для решений большинства практических задач многомерного статистического анализа методами теории одномерных случайных величин.

Плотность распределения $f(X)$, $X \in R^p$, называется радиальной, если $f(X) = c_p f_1(||X||)$, где $f_1(y)$ — одномерное симметричное распределение. Из свойств проекции распределения следует, что $f(X)$ — радиальное распределение тогда и только тогда, когда $\breve{f}(y, \mathbf{a}_1) \equiv \breve{f}(y, \mathbf{a}_2)$ для любых единичных векторов $\mathbf{a}_1$ и $\mathbf{a}_2$. Заметим, что $f_1(y)$ является плотностью распределения случайной величины y, задаваемой ограничением радиального случайного вектора X на прямую $X = yX_0$ для некоторого фиксированного X_0, $||X_0|| = 1$. Далее будем рассматривать только ортогональные проекции, поэтому для радиальных распределений можно положить $\breve{f}(y, \mathbf{a}) \equiv \breve{f}(y)$. Важные примеры радиального распределения дают p-мерное нормальное распределение $N(X; 0, \sigma^2\mathbf{I}_p) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{p-1} f_1(||X||)$, где $f_1(y) = N(y; 0, \sigma^2)$, и равномерное распределение $R(X; 0, r^2\mathbf{I}_p)$ в шаре $D^p \subset R^p$ с центром в начале координат и радиуса r, где $f_1(y) = \theta(y; r)$ и $\theta(y; r) = \frac{1}{2r}$, если $|y| \leqslant r$, и $\theta(y, r) = 0$, когда $|y| > r$.

Лемма 20.1. Формула $f(X) = c_p f_1(||X||)$ задает радиальное распределение в R^p тогда и только тогда, когда $f_1(y)$ — одномерное симметричное распределение с конечным $(p-1)$-м центральным моментом ν_{p-1} и $c_p = \Gamma\left(\frac{p}{2}\right) \Big/ \pi^{p/2}\nu_{p-1}$.

Заметим, что ковариационная матрица радиального распределения $f(X)$ есть $\sigma^2\mathbf{I}_p$, где $\sigma^2 = \frac{1}{p}\int ||X||^2 f(X)\,dX$. Согласно формуле (20.3) для любого невырожденного преобразования $\mathbf{Q}: R^p \to R^p$ и радиального распределения $f_\xi(X) = c_p f_1(||X||)$ имеет место формула

$$f_{\mathbf{Q}\xi + x_0}(X) = c_p |\det \Sigma|^{-1/2} f_1(((X - X_0)' \Sigma^{-1} (X - X_0))^{1/2}),$$

где $\Sigma = \mathbf{Q}_p\mathbf{Q}_p'$ — ковариационная матрица случайного вектора $\mathbf{Q}\xi$.

Лемма 20.2. Одномерная проекция $\breve{f}$ радиальной плотности $f(X) = c_p f_1(\|X\|)$ задается формулой

$$\breve{f}(y) = \frac{2}{\nu_{p-1}\beta\left(\frac{1}{2}, \frac{p-1}{2}\right)} \int_0^\infty f_1\left((t^2+y^2)^{1/2}\right) t^{p-2}\,dt, \qquad (20.18)$$

или

$$\breve{f}(y) = \frac{1}{\nu_{p-1}\beta\left(\frac{1}{2}, \frac{p-1}{2}\right)} \int_0^\infty (\tau^2 - y^2)_+^{\frac{p-1}{2}-1} f_1(\tau)\,d\tau^2, \qquad (20.19)$$

где $\beta\left(\frac{1}{2}, \frac{p-1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p-1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}$ — бета-функция.

Пример 20.2. Пусть $f(X) = N(X, 0, \sigma^2 \mathbf{I}_p)$. Тогда $f(X) = c_p f_1(\|X\|)$, где $f_1(y) = N(y, 0, \sigma^2)$, и $c_p = \frac{\Gamma(p/2)}{\pi^{p/2}\nu_{p-1}}$ с $\nu_{p-1} = (\sqrt{2\pi}\sigma)^{p-1}\Gamma\left(\frac{p}{2}\right)$. Следовательно,

$$\breve{f}(y) = N(y, 0, \sigma^2). \qquad (20.20)$$

Пример 20.3. Рассмотрим случай равномерного распределения $R(X, 0, r^2\mathbf{I}_p)$ в шаре $D^p \subset R^p$ радиуса r. Имеем $f(X) = c_p f_1(\|X\|)$, где $f_1(y) = \theta(y, r) = \frac{1}{2r}(r^2 - y^2)_+^0$ и $\nu_{p-1} = \frac{r^{(p-1)}}{p}$. Следовательно,

$$\breve{f}(y) = \frac{p}{(p-1)} \frac{1}{\beta\left(\frac{1}{2}, \frac{p-1}{2}\right) r} \left(1 - \frac{y^2}{r^2}\right)_+^{\frac{p-1}{2}}. \qquad (20.21)$$

20.2.2. Важные модели радиальных распределений. Механизмы формирования случайных векторов с модельными радиальными распределениями. В качестве моделей распределений, сосредоточенных в шаре $D^p \subset R^p$, удобно использовать представителей семейства распределений, описываемого следующей теоремой.

Т е о р е м а 20.1. Для каждого p формула

$$R_{p,\alpha}(X;\ 0,\ \sigma^2\mathbf{I}_p)=\frac{\Gamma\left(\frac{p}{2}+\alpha\right)}{[(2\alpha+p)\pi]^{p/2}\Gamma(\alpha)\sigma^p}\times$$

$$\times\left(1-\frac{\|X\|^2}{(2\alpha+p)\sigma^2}\right)_+^{\alpha-1} \qquad (20.22)$$

задает двупараметрическое семейство (по $\alpha>0$ и σ) радиальных распределений, сосредоточенных в шаре радиуса $\sqrt{2\alpha+p}\sigma$, где σ^2 — дисперсия

Одномерная проекция распределения $R_{p,\alpha}$ имеет вид:

$$\widetilde{R}_{p,\alpha}=R_{1,\ \alpha+\frac{p-1}{2}}(y,\ 0,\ \sigma^2). \qquad (20.23)$$

Заметим, что $R_{p,1}(X,\ 0,\ \sigma^2\mathbf{I}_p)$ представляет собой равномерное распределение в шаре радиуса $\sqrt{2+p}\,\sigma$, а при фиксированной дисперсии σ^2 и $\alpha\to\infty$ распределение $R_{p,\alpha}$ переходит в p-мерное нормальное распределение Таким образом, формула (20 23) в качестве частных случаев содержит формулы (20 20) и (20 21) Она показывает, что семейство $R_{p,\alpha}(X,\ 0,\ \sigma^2\mathbf{I}_p)$ при фиксированном σ^2 замкнуто относительно оператора проецирования, который на этом семействе в явном виде показывает свои сглаживающие свойства. при натуральном α и нечетном p он переводит $(\alpha-1)$ раз дифференцируемую функцию в функцию дифференцируемую $(\alpha-1)+\frac{p-1}{2}$ раз

Отметим, что и в случае общего p-мерного распределения $f_\xi(X)$ необходимо учитывать это свойство оператора проецирования при подборе модели одномерного распределения $\widetilde{f}_\xi(y,\ a)$, если из каких-либо соображений уже выбран класс гладкости модели p-мерного распределения $f_\xi(x)$.

Опишем схему (механизм) формирования случайных векторов с плотностью распределения $R_{p,\alpha}(X;\ 0,\ \sigma^2\mathbf{I}_p)$ для $\alpha=l/2$, где l — натуральное число.

Пусть $\eta=(\eta^1,\ \ldots,\ \eta^p)$ и $\zeta=(\zeta^1,\ \ldots,\ \zeta^l)$ — случайные независимые векторы, распределенные по нормальным законам $N(0,\ \mathbf{I}_p)$ и $N(0,\ \mathbf{I}_l)$ соответственно.

Положим $\eta\times\zeta=(\eta^1,\ \ldots,\ \eta^p,\ \zeta^1,\ \ldots\ \zeta^l)$ и

$$\xi_{p,l}=\frac{\eta}{\|\eta\times\zeta\|},$$

где $\|\eta\times\zeta\|=(\|\eta\|^2+\|\zeta\|^2)^{1/2}$.

Ясно, что вектор $\xi_{p,\,l} \in R^p$ распределен по радиальному закону, поэтому для вычисления закона распределения его одномерной проекции достаточно вычислить проекцию на одну из координатных осей, скажем $\mathbf{e}_1$. Имеем

$$\xi'_{p,\,l}\mathbf{e} = \frac{\eta^1}{\|\eta \times \zeta\|}$$

Известно, что случайная величина

$$\frac{\theta^1}{\sqrt{\frac{1}{m}\sum_1^m (\theta^\nu)^2}}, \text{ где } \theta = (\theta^1, \ldots, \theta^m) \sim N(0, \mathbf{I}_m),$$

распределена по закону $\dfrac{1}{\sqrt{m\pi}} \dfrac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m-1}{2}\right)} \left(1 - \frac{y^2}{m}\right)_+^{\frac{m-3}{2}}$,

$m > 1$.

Так как $\eta \times \zeta = (\eta^1, \ldots, \eta^p, \zeta^1, \ldots, \zeta^l) \sim N(0, \mathbf{I}_{p+l})$, то получаем, что случайная величина $\xi'_{p,\,l}\mathbf{e}_1$ распределена по закону

$$\frac{1}{\beta\left(\frac{1}{2}, \frac{p+l-1}{2}\right)} (1-y^2)_+^{\frac{p+l-1}{2}-1},$$

т. е. $r^2 = 1 = (p+l)\sigma^2$ и $f_{\xi'_{p,\,l}\mathbf{e}_1}(y) = R_{1,\,\frac{p+l-1}{2}}\left(y;\ 0, \frac{1}{p+l}\right)$.

Следовательно, для любого $\lambda > 0$

$$f_{\lambda(\xi'_{p,\,l}\mathbf{e}_1)} = R_{1,\,\frac{p+l-1}{2}}\left(y;\ 0, \frac{\lambda^2}{p+l}\right)$$

Используя теперь, что многомерное распределение полностью определяется своими одномерными проекциями, из формулы (20.22) получаем:

$$f_{\lambda\xi_{p,\,l}}(X) = R_{p,\,\frac{l}{2}}(X;\ 0,\ \sigma^2\mathbf{I}_p),$$

где $\lambda = \sqrt{l+p}\,\sigma$. В частности, согласно формуле (20.22) случайный вектор $\sqrt{2+p}\sigma\xi_{p,\,2}$ имеет равномерное распределение в шаре радиуса $\sqrt{2+p}\,\sigma$. Для $\alpha > 0$ распределение $R_{p,\,\alpha}(X;\ 0,\ \sigma^2\mathbf{I}_p)$ является невырожденным

Заметим, что при $\alpha \to +0$ распределение $R_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p)$ стремится к вырожденному распределению, которое представляет собой равномерное распределение, сосредоточенное на сфере радиуса σ. Это распределение имеет случайный вектор $\sigma\xi_{p,0} = \frac{\sigma\eta}{\|\eta\|}$.

Опишем теперь радиальные распределения, связанные с распределением Стьюдента. Имеет место следующая теорема.

Т е о р е м а 20.2. Для каждого p формула

$$t_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p) = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{[\pi(\alpha-p+1)]^{p/2}\,\Gamma\left(\frac{\alpha+1-p)}{2}\right)\sigma^p} \times$$

$$\times \left(1 + \frac{\|X\|^2}{(\alpha-p+1)\,\sigma^2}\right)^{-\frac{\alpha+1}{2}} \qquad (20.24)$$

задает двупараметрическое семейство (по $\alpha > p-1$ и σ) радиальных распределений, где σ^2 — дисперсия.

Одномерная проекция распределения $t_{p,\alpha}$ имеет вид:

$$\widetilde{t}_{p,\alpha} = t_{1,\alpha-p+1}(y, 0, \sigma^2). \qquad (20.25)$$

Заметим, что $t_{1,1}(X; 0, \sigma^2)$ представляет собой распределение Коши. Из условия $\alpha > p-1$ следует, что не существует радиальных распределений $t_{p,1}(X; 0, \sigma^2\mathbf{I}_p)$ при $p > 2$. Для натуральных m $t_{1,m}(X; 0, \sigma^2)$ задает распределение Стьюдента с m степенями свободы и для всех $p < m+1$ существует радиальное распределение $t_{p,m}(X; 0, \sigma^2\mathbf{I}_p)$. Отметим также, что $t_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p) \to N(X; 0, \sigma^2\mathbf{I}_p)$ при $\alpha \to \infty$ и фиксированной дисперсии σ^2.

Опишем схему формирования случайных векторов с плотностью распределения $t_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p)$ для целых α.

Известно, что случайная величина

$$\frac{\theta^1}{\sqrt{\frac{1}{m}\sum_{2}^{m+1}(\theta^\nu)^2}}, \text{ где } (\theta^1, \ldots, \theta^{m+1}) \sim N(0, \mathbf{I}_{m+1})$$

имеет распределение Стьюдента с m степенями свободы

$$t_m(y) = \frac{1}{\sqrt{m}\,\beta\left(\frac{1}{2}, \frac{m}{2}\right)}\left(1 + \frac{y^2}{m}\right)^{-\frac{m+1}{2}}.$$

Пусть $\eta = (\eta^1, \ldots, \eta^p)$ и $\zeta = (\zeta^1, \ldots, \zeta^m)$ — случайные независимые векторы, распределенные по нормальным законам $N(0, \mathbf{I}_p)$ и $N(0, \mathbf{I}_m)$ соответственно.

Рассмотрим p-мерный случайный вектор $\xi = \frac{\eta}{\|\zeta\|}$. Он распределен по радиальному закону, а, как следует из механизма формирования закона Стьюдента, его одномерная проекция распределена по закону

$$f_{1,m}(y) = \frac{1}{\sqrt{m}\beta\left(\frac{1}{2}, \frac{m}{2}\right)}\left(1 + \frac{y^2}{m}\right)^{-\frac{m+1}{2}}.$$

Используя теперь формулу (20.25), получаем, что p-мерный случайный вектор $r\xi$ распределен по закону $f_{p, m+p-1}(X)$. В качестве следствия получаем:
если $(\eta^1, \ldots, \eta^p, \eta^{p+1}) \sim N(0, \mathbf{I}_{p+1})$, то p-мерный вектор $\xi = r\frac{\eta}{|\eta^{p+1}|}$, где $\eta = (\eta^1, \ldots, \eta^p)$ распределен по закону

$$f_{pp}(X) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\pi^{\frac{p+1}{2}} r^p}\left(1 + \frac{\|X\|^2}{r^2}\right)^{-\frac{p+1}{2}},$$

а одномерная его проекция распределена по закону Коши:

$$f_{1,1}(y) = \frac{1}{\pi r}\left(1 + \frac{y^2}{r^2}\right)^{-1}.$$

20.2.3. Экстремальные многомерные распределения. Пусть $F_p(X_0, \Sigma)$ — совокупность всех плотностей p-мерных случайных векторов ξ с фиксированными вектором средних X_0 и невырожденной ковариационной матрицей Σ.

Известно, (см. например, [129, с. 476]), что интеграл энтропии

$$H(f_\xi) = -\int f_\xi(X) \log f_\xi(X)\, dX$$

для $f_\xi \in F_p(X_0, \Sigma)$ достигает максимального значения, только когда

$$f_\xi(X) = N(X;\ X_0,\ \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}} |\det \Sigma|^{\frac{1}{2}}} \times$$

$$\times e^{-\frac{1}{2}(X - X_0)' \Sigma^{-1} (X - X_0)},$$

причем $H(N(X;\ X_0,\ \Sigma)) = \frac{1}{2}(p + p \log(2\pi) + \log|\det \Sigma|)$.

Покажем, что аналогичный результат имеет место для рассмотренного выше семейства плотностей

$$R_{p,\alpha}(X;\ X_0,\ \Sigma) = c_{p,\alpha}\left(1 - \frac{1}{2\alpha+p}(X-X_0)'\,\Sigma^{-1}\times\right.$$

$$\left.\times\,(X-X_0)\right)_+^{\alpha-1},\ \alpha>1,$$

где $c_{p,\alpha} = \dfrac{1}{(2\pi)^{\frac{p}{2}}\,|\det\Sigma|^{\frac{1}{2}}}\,\dfrac{\Gamma\left(\frac{p}{2}+\alpha\right)}{\left(\frac{p}{2}+\alpha\right)^{\frac{p}{2}}\,\Gamma(\alpha)}.$

Пусть $\tau = \varphi(t)$ — некоторая непрерывная строго возрастающая функция при $t \geqslant 0$ и $\varphi(0) = 0$. Тогда для любых $a \geqslant 0$ и $b \geqslant 0$ имеет место *неравенство Юнга*

$$ab \leqslant \int_0^a \varphi(t)\,dt + \int_0^b \varphi^{-1}(\tau)\,d\tau, \tag{20.26}$$

где $\varphi^{-1}(\varphi(t)) = t$, причем равенство достигается только при $b = \varphi(a)$. Рассмотрим случай, когда $\tau = t^{\frac{1}{\alpha-1}}$, где $\alpha > 1$. Тогда $\varphi^{-1}(\tau) = t = \tau^{\alpha-1}$ и из (20.26) получаем: для любых $a \geqslant 0$ и $b \geqslant 0$ имеет место неравенство

$$ab \leqslant \left(1-\frac{1}{\alpha}\right)a a^{\frac{1}{\alpha-1}} + \frac{1}{\alpha}\,b b^{\alpha-1}, \tag{20.27}$$

причем равенство достигается только при $b = a^{\frac{1}{\alpha-1}}$.

Положим $a = a_1$ и $b = a_2^{\frac{1}{\alpha-1}}$. Тогда из (20.27) получаем неравенство

$$a_1\,a_2^{\frac{1}{\alpha-1}} \leqslant \left(1-\frac{1}{\alpha}\right)a_1\,a_1^{\frac{1}{\alpha-1}} + \frac{1}{\alpha}\,a_2\,a_2^{\frac{1}{\alpha-1}}. \tag{20.28}$$

Ясно, что (20.28) эквивалентно неравенству

$$a_1\left[\alpha\left(a_2^{\frac{1}{\alpha-1}}-1\right)\right] \leqslant a_1\left[\alpha\left(a_1^{\frac{1}{\alpha-1}}-1\right)\right] -$$

$$-a_1\,a_1^{\frac{1}{\alpha-1}} + a_2\,a_2^{\frac{1}{\alpha-1}}. \tag{20.29}$$

Используя теперь, что $\alpha\,(t^{\frac{1}{\alpha-1}}-1) \to \ln t$ при $\alpha \to \infty$, получаем из (20.29) неравенство $a_1 \ln a_2 \leqslant a_1 \ln a_1 - a_1 + a_2$,

лежащее в основе доказательства экстремальности многомерного нормального закона.

Положим

$$H_\alpha(f_\xi) = -\int f_\xi(X)\left[\alpha\left(f_\xi(X)^{\frac{1}{\alpha-1}} - 1\right)\right] dX,$$

$$h_\alpha(f_\xi) = -\int f_\xi(X) f_\xi(X)^{\frac{1}{\alpha-1}} dX.$$

Например, для целых $m \geqslant 2$

$$h_{\frac{m}{m-1}}(f_\xi) = -\int f_\xi^m(X)\, dX = -\|f_\xi(X)\|_m^m,$$

в частности, $h_2(f_\xi) = -\|f_\xi(X)\|_2^2$.

Заметим, что $H_\alpha(f_\xi) \to H(f_\xi)$ при $\alpha \to \infty$. Так как $H_\alpha(f_\xi) = \alpha(h_\alpha(f_\xi) + 1)$, то для каждого α, $1 < \alpha < \infty$, задачи на экстремум для функционалов $H_\alpha(f_\xi)$ и $h_\alpha(f_\xi)$ эквивалентны.

Полагая в (20.29) $a_1 = f_\xi(X)$, $a_2 = R_{p,\alpha}(X; X_0, \Sigma)$ и интегрируя его по R^p, получаем после умножения на $-\frac{\alpha}{\alpha-1}$, что имеет место следующая теорема.

Теорема 20.3. Функционалы $h_\alpha(f_\xi)$ и $H_\alpha(f_\xi)$ достигают максимальные значения, только когда $f_\xi(X) = R_{p,\alpha}(X; X_0, \Sigma)$, причем

$$h_\alpha(R_{p,\alpha}) = -\frac{2\alpha}{2\alpha+p} c_{p,\alpha}^{\frac{1}{\alpha-1}};$$

$$H_\alpha(f_\xi) = \alpha\left(1 - \left[\frac{c_{p,\alpha}}{\left(1+\frac{p}{2\alpha}\right)^{\alpha-1}}\right]^{\frac{1}{\alpha-1}}\right).$$

Положим

$$h_\alpha^*(f_\xi) = |\det 2\pi\Sigma|^{\frac{1}{2(\alpha-1)}} h_\alpha(f_\xi).$$

Заметим, что для любого невырожденного линейного преобразования $\mathbf{Q}: R^p \to R^p$:

$$h_\alpha^*(f_{\mathbf{Q}\xi+X_0}) = h_\alpha^*(f_\xi).$$

Следствие 20.1. Функционал h_α^* на множестве всех плотностей p-мерных случайных векторов с невырожденной ковариационной матрицей является инвариантным относительно невырожденных линейных преобразований, ограни-

чен сверху константой c и достигает максимальное значение на радиальной плотности $R_{p,\alpha}$,

где $c = \dfrac{\alpha}{\frac{p}{2}+\alpha} \dfrac{\Gamma\left(\frac{p}{2}+\alpha\right)}{\Gamma(\alpha)\left(\frac{p}{2}+\alpha\right)^{\frac{p}{2}}}$.

20.3. Теория процедур оптимизации проекционных индексов

Пусть $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ — выборка объема n в R^p. Каждая статистика $\varphi(Y_1, \ldots, Y_n)$ на q-мерных выборках $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)$ объема n, $1 \leqslant q < p$, задает проекционный индекс $F(\mathbf{A}) = \varphi(\mathbf{A}X_1, \ldots, \mathbf{A}X_n)$, где $\mathbf{A}$ — некоторая проекция из R^p в R^q.

Обозначим через $ML(p, q)$ — множество всех проекций из R^p в R^q (см. п. 20.2.1). Решение статистической задачи методом целенаправленного проецирования содержит два этапа:

1) выбор проекционного индекса $F(\mathbf{A})$;

2) решение оптимизационных задач для функции $F(\mathbf{A})$ на соответствующем подмножестве $M \subseteq ML(p, q)$.

Вопросы, связанные с этапом 1, достаточно подробно разобраны в гл. 19. Основная цель настоящего параграфа — изложить теорию алгоритмов решения задач этапа 2.

20.3.1. Области оптимизации в задачах поиска выразительных проекций. Начнем со структуры множества $ML(p, q)$. Как отмечалось в п. 20.2.1, каждая проекция из R^p в R^q однозначно определяется $(p \times q)$-матрицей $\mathbf{A}$, удовлетворяющей условию $\det \mathbf{AA}' \neq 0$. Поставим в соответствие $(p \times q)$-матрице $\mathbf{A}$ набор из q ее p-мерных вектор-строк $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$ и заметим, что $\det \mathbf{AA}' \neq 0$ тогда и только тогда, когда эти векторы линейно независимы.

Пусть $\mathbf{L}(\mathbf{A})$ — линейное q-мерное подпространство в R^p, натянутое на векторы $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q\}$, составляющие матрицу проекции $\mathbf{A}$. Тогда имеет место следующая лемма

Лемма 20.3. Соответствие $\mathbf{A} \to (\mathbf{L}(\mathbf{A}), \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q\})$ позволяет отождествить $ML(p, q)$ с множеством всех пар $(\mathbf{L}, (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q))$, где $\mathbf{L}$ — q-мерное подпространство в R^p, а $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q\}$ — некоторый базис в этом подпространстве.

Аналогично пусть $MO(p, q)$ — множество всех ортогональных проекций из R^p в R^q (см. п. 20.2.1). Тогда верна следующая лемма.

Лемма 20.4. Соответствие $\mathbf{A} \to (L(\mathbf{A}), \{\theta_1, \ldots, \theta_q\})$ позволяет отождествить $O(p, q)$ с множеством всех пар $(L, \{\theta_1, \ldots, \theta_q\})$, где L — q-мерное подпространство в R^p, а $\{\theta_1, \ldots, \theta_q\}$ — некоторый *ортонормированный базис* в этом подпространстве.

Указанное соответствие отождествляет, в частности, множество всех проекций из R^p в R^1 с множеством ненулевых векторов $\theta \in R^p$, а множество всех ортогональных проекций из R^p в R^1 с множеством единичных векторов θ (см. п. 20.2.1).

Определение 20.1. Статистика $\varphi(Y_1, \ldots, Y_n)$ на q-мерных выборках называется инвариантной относительно преобразования $\mathbf{B}$ пространства R^q, если $\varphi(\mathbf{B}Y_1, \ldots, \mathbf{B}Y_n) = \varphi(Y_1, \ldots, Y_n)$.

Практически все важные проекционные индексы строятся по статистикам, инвариантным относительно невырожденных либо ортогональных преобразований $\mathbf{B}$.

Определение 20.2. Проекционный индекс F, построенный по статистике $\varphi(Y_1, \ldots, Y_n)$, инвариантной относительно *любого невырожденного* преобразования, называется GL-инвариантным и соответственно O-инвариантным, если $\varphi(Y_1, \ldots, Y_n)$ инвариантна относительно *любого ортогонального* преобразования.

Введем теперь так называемое *многообразие Грассмана* $G(p, q)$, точками которого являются q-мерные линейные подпространства $\mathbf{L}$ в R^p [118][1].

Возьмем некоторый GL-инвариантный проекционный индекс $F(\mathbf{A})$. Рассмотрим q-мерное линейное подпространство $\mathbf{L} \subset R^p$, выберем в нем базис $\theta_1, \ldots, \theta_q$ и тем самым получим проекцию $\mathbf{A}(\mathbf{L})$ из R^p в R^q. Положим, по определению,

$$\Phi(\mathbf{L}) = F(\mathbf{A}(\mathbf{L})). \tag{20.30}$$

Проекция $\mathbf{A}(\mathbf{L})$ определена с точностью до выбора базиса в $\mathbf{L}$, но так как $F(\mathbf{A})$ является GL-инвариантным, то формула (20.30) корректно задает функцию на многообразии Грассмана $G(p, q)$.

---

[1] Многообразие $G(p, q)$ — классический математический объект, названо в честь немецкого математика, физика и филолога Г. Грассмана (1809—1877). В сочинении «Учение о линейном пространстве» Г. Грассман дал первое систематическое построение теории многомерного евклидова пространства, ввел скалярное произведение векторов [30].

Следствие 20.2. Задача поиска выразительной проекции $\mathbf{A}_*$ для GL-инвариантного проекционного индекса $F(\mathbf{A})$ эквивалентна задаче:

$$\text{найти } \mathbf{L}_* = \arg \underset{\mathbf{L} \in G(p, q)}{\text{extr}} \Phi(\mathbf{L}). \tag{20.31}$$

Следствие 20.3. Задача поиска выразительной ортогональной проекции $\mathbf{A}_*$ для O-инвариантного проекционного индекса $F(\mathbf{A})$ также эквивалентна задаче (20.31). Естественно, при построении $\mathbf{A}_* = \mathbf{A}(\mathbf{L}_*)$, в этом случае необходимо взять некоторый ортогональный базис в найденном экстремальном подпространстве $\mathbf{L}_* \subset \mathbf{R}^p$.

Таким образом, показано, что поиск выразительных проекций из R^p в R^q сводится к решению оптимизационных задач на многообразии Грассмана $G(p, q)$.

Объясним, какие преимущества дает эта редукция. Отождествляя, как обычно, пространство всех $(p \times q)$-матриц $\mathbf{A} = (a_{ij},\ 1 \leqslant i \leqslant p,\ 1 \leqslant j \leqslant q)$ с евклидовым пространством R^{pq}, получаем, что $ML(p, q)$ является открытой областью в R^{pq}, выделяемой условием $\det |\mathbf{A}\mathbf{A}'| \neq 0$, а $MO(p, q)$ — замкнутым подмногообразием в R^{pq}, выделяемым условием $\mathbf{A}\mathbf{A}' = \mathbf{I}_q$, т. е. $q(q+1)/2$ уравнениями, связывающими pq координат $\{a_{ij}\}$.

Например, $MO(p, 1) = \{\theta \in R^p,\ \|\theta\| = 1\}$ (одно уравнение связи на p координат);

$$MO(p, 2) = \{(\theta_1, \theta_2) \in R^p \times R^p = R^{2p},\ \|\theta_1\| = 1,\ \|\theta_2\| = 1,\ \theta_1' \theta_2 = 0)\}$$

(три уравнения связи на $2p$ координат).

Размерность многообразия $MO(p, q)$ на $\frac{q(q+1)}{2}$ меньше, чем размерность объемлющего пространства R^{pq}. Таким образом, применение обычных численных процедур для решения оптимизационных задач на многообразиях $ML(p, q)$ и $MO(p, q)$ как подмножествах евклидова пространства R^{pq} практически невозможно. Отметим, что попытка построить метод градиентного спуска на $MO(p, q)$ была предпринята в [121, § 6.10]. Трудности, которые встретились на этом пути, типичны для реализации процедур *условной оптимизации* при большом числе уравнений связи на координаты.

В [37—39] разработаны численные методы оптимизации функций на многообразии $G(p, q)$. Оказалось, что если использовать внутреннюю геометрию этого многообразия, то эти методы реализуются при помощи аналогов основных алгоритмов *безусловной оптимизации*.

20.3.2. Алгоритмы оптимизации функций на многообразиях проекций. Пусть $\Phi(\mathrm{L})$ — некоторая функция на многообразии $G(p, q)$. Опишем сначала алгоритм решения задачи (20.31), реализующий аналог методов покоординатной оптимизации. Фиксируем в R^p некоторый ортонормированный базис $\Theta_0 = \{\theta_{10}, \ldots, \theta_{p0}\}$ и обозначим через $\mathrm{L}_0 = \mathrm{L}(\Theta)$ подпространство, натянутое на первые q векторов $\{\theta_1, \ldots, \ldots, \theta_q\}$. Возьмем L_0 за начальную точку алгоритма оптимизации. Допустим, что уже построены $\mathrm{L}_0, \ldots, \mathrm{L}_m$, причем каждое L_m представляет собой q-мерное подпространство в R^p, натянутое на первые q векторов ортонормированного базиса $\Theta_m = \{\theta_{1m}, \ldots, \theta_{pm}\}$, т. е. $\mathrm{L}_m = \mathrm{L}(\Theta_m)$.

Опишем переход от (Θ_m, L_m) к $(\Theta_{m+1}, \mathrm{L}_{m+1})$. Выберем i, $1 \leqslant i \leqslant q$ и j, $1 \leqslant j \leqslant p - q$, введем семейство базисов $\Theta_m(i, j, \alpha) = \{\theta_{1m}(i, j), \ldots, \theta_{pm}(i, j)\}$, где

$$\theta_{im}(ij) = \cos\alpha\theta_{im} + \sin\alpha\theta_{i+j,m},\ \theta_{i+j,m}(i, j) =$$
$$= -\sin\alpha\theta_{im} + \cos\alpha\theta_{i+j,m}, \qquad (20.32)$$
$$\theta_{lm}(i, j) = \theta_{lm},\ \text{если}\ l \neq i\ \text{или}\ i+j,$$

и получим семейство q-мерных плоскостей $\mathrm{L}_m(i, j; \alpha)$, где $\mathrm{L}_m(i, j, \alpha)$ натянуто на $\{\theta_{1m}(i, j), \ldots, \theta_{qm}(i, j)\}$.

Определение 20.3. Семейство q-мерных плоскостей $\mathrm{L}_m(i, j; \alpha) \subset R^p$ называется (i, j)-координатной линией в $G(q, p)$, проходящей через точку $\mathrm{L}_m = \mathrm{L}_m(i, j; 0)$ в локальной системе координат, задаваемой базисом Θ_m.

Обоснование такого определения см. в [37], где описаны все необходимые факты о структуре многообразия $G(p, q)$.

Ограничив функцию $\Phi(\mathrm{L})$ на семейство плоскостей $\mathrm{L}_m(i, j; \alpha)$ получаем обычную числовую функцию от α, $\alpha \in [0, 2\pi]$. Положим $\Phi(\mathrm{L}_m(i, j; \alpha)) = \Phi_{ij}(\alpha)$. Используя одну из стандартных процедур оптимизации числовой функции $\Phi_{ij}(\alpha)$, находим

$$\alpha_* = \arg \max_{\alpha \in [0, 2\pi]} \Phi_{ij}(\alpha)$$

и полагаем $\Theta_{m+1} = \Theta(i, j; \alpha_*)$, $\mathrm{L}_{m+1} = \mathrm{L}_m(i, j; \alpha_*)$. Цикл процедуры оптимизации завершен.

Используя описанный шаг алгоритма, можно реализовать методы покоординатной оптимизации. Например, упорядочив множество пар (i, j), $1 \leqslant i \leqslant q$, $1 \leqslant j \leqslant p - q$ и зациклив его, можно получить сколь угодно длинную последовательность пар (i, j) и соответствующую последовательность точек $\mathrm{L}_0, \mathrm{L}_1, \ldots, \mathrm{L}_m, \mathrm{L}_{m+1}, \ldots$ Заметим, что по построению,

$$\Phi(\mathrm{L}_0) < \Phi(\mathrm{L}_1) < \ldots < \Phi(\mathrm{L}_m) < \Phi(\mathrm{L}_{m+1}) < \ldots$$

Если функция $\Phi(\mathrm{L})$ непрерывно зависит от L, то, используя компактность многообразия $G(p, q)$, получаем, что $\Phi(\mathrm{L})$ — ограниченная сверху функция и последовательность $\{\Phi(\mathrm{L}_m)\}$ сходится.

Рассмотрим теперь дифференцируемую функцию $\Phi(\mathrm{L})$ на $G(p, q)$. Опишем алгоритмы, реализующие методы градиентной оптимизации.

Пусть, как и выше, $\mathrm{L}_0 = \mathrm{L}(\boldsymbol{\Theta}_0) \subset R^p$, где $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{p0}\}$ — ортогональный базис в R^p и $\mathrm{L}_0(i, j; \alpha)$ — (i, j)-я координатная линия в $G(p, q)$, проходящая через L_0.

Определение 20.4. Градиентом функции $\Phi(\mathrm{L})$ в точке L_0 относительно локальной системы координат, задаваемой базисом $\boldsymbol{\Theta}_0$, называется $q(p-q)$-мерный вектор $\boldsymbol{\Gamma} = \{\Gamma_{ij},\ 1 \leqslant i \leqslant q,\ 1 \leqslant j \leqslant p-q\}$, вычисляемый по формуле

$$\Gamma_{ij} = \frac{\mathrm{d}}{\mathrm{d}\alpha} \Phi_{ij}(\alpha)\,|_{\alpha=0}, \tag{20.33}$$

где $\Phi_{ij}(\alpha) = \Phi(\mathrm{L}_0(i, j; \alpha))$.

Градиент $\boldsymbol{\Gamma} = \{\Gamma_{ij}\}$ удобно представлять в виде $q \times (p-q)$-матрицы с вектор-столбцами $\boldsymbol{\gamma}_i = (\gamma_{ij})$, $1 \leqslant i \leqslant q$.

Пример 20.4. Пусть $X \in R^p$ и X_{L} — его ортогональная проекция на некоторое q-мерное линейное подпространство $\mathrm{L} \subset R^p$. Рассмотрим на многообразии $G(p, q)$ функцию $\Phi(\mathrm{L}) = ||X_{\mathrm{L}}||^2$ и вычислим ее градиент в точке $\mathrm{L}_0 = \mathrm{L}(\boldsymbol{\Theta}_0)$.

Из (20.31) следует

$$\Phi(\mathrm{L}_0(i, j; \alpha)) = \sum_{\substack{l=1 \\ l \neq i}}^{q} (\boldsymbol{\theta}'_{l0} X)^2 + (\cos\alpha \boldsymbol{\theta}'_{i0} X + \sin\alpha \boldsymbol{\theta}'_{i+j,0} X)^2, \tag{20.34}$$

поэтому непосредственно из (20.33) получаем:

$$\Gamma_{ij} = 2(\boldsymbol{\theta}'_{i0} X)(\boldsymbol{\theta}'_{i+j} X) = 2\boldsymbol{\theta}'_{i0}(XX')\boldsymbol{\theta}_{i+j},\ 1 \leqslant i \leqslant q,\ 1 \leqslant j \leqslant p-q. \tag{20.35}$$

Пусть $\mathbf{A}(\mathrm{L}_0)$ — проекция из R^p в R^q, задаваемая $(p \times q)$-матрицей, строки которой $\boldsymbol{\theta}'_{10}, \ldots, \boldsymbol{\theta}'_{q0}$, и $\mathbf{A}(\mathrm{L}_0^{\perp})$ — проекция из R^p в R^{p-q}, задаваемая $(p \times (p-q))$-матрицей со строками $\boldsymbol{\theta}'_{q+1,0}, \ldots, \boldsymbol{\theta}'_{p,0}$. Тогда из (20.35) следует, что градиент функции $||X_{\mathrm{L}}||^2$ в точке L_0 относительно базиса $\boldsymbol{\Theta}_0$ можно записать в виде $q \times (p-q)$-матрицы

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\boldsymbol{\Theta}_0} = 2\mathbf{A}(\mathrm{L}_0) XX' \mathbf{A}(\mathrm{L}_0^{\perp})'. \tag{20.36}$$

При помощи функции $||X_{\mathbf{L}}||^2$ легко показать, как зависит вид градиента $\Gamma = \Gamma_{\Theta}$ от выбора базиса Θ.

Пусть $\Theta_k = \{\theta_{1k}, \ldots, \theta_{pk}\}$, $k = 1, 2$ — ортонормированные базисы в R^p.

Заметим, что $\mathbf{L}(\Theta_1) = \mathbf{L}(\Theta_2) = \mathbf{L}_0$ тогда и только тогда, когда существуют ортогональная $(q \times q)$-матрица $\mathbf{B}$, переводящая базис $\{\theta_{11}, \ldots, \theta_{q1}\}$ в подпространстве $\mathbf{L}_0 \subset R^p$ в его базис $\{\theta_{12}, \ldots, \theta_{q2}\}$, и ортогональная $(p-q) \times (p-q)$-матрица $\mathbf{B}^{\perp}$, переводящая базис $\{\theta_{q+1,1}, \ldots, \theta_{p,1}\}$ в ортогональном дополнении $\mathbf{L}_0^{\perp}$ к $\mathbf{L}_0 \subset R^p$ в его базис $\{\theta_{q+1,2}, \ldots, \theta_{p,2}\}$.

Следовательно, если $\mathbf{L}(\Theta_1) = \mathbf{L}(\Theta_2) = \mathbf{L}_0$, то

$$\Gamma_{\Theta_2} = \mathbf{B}' \Gamma_{\Theta_1} \mathbf{B}^{\perp}.$$

Пример 20.5. Пусть $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ — выборка в R^p. Вычислим градиент функции $\Phi(\mathbf{L})$ на многообразии Грассмана, соответствующей проекционному индексу $F(\mathbf{A})$ для статистики $\varphi(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} ||Y_i - Y_0||^2$, где Y_0 — средний вектор выборки $Y^{(n)}$.

Имеем $\Phi(\mathbf{L}) = \frac{1}{n} \sum_{i=1}^{n} \| (X_i - X_0)_L \|^2$.

Используя формулу (20.33), получаем в точке $\mathbf{L}_0 = \mathbf{L}(\Theta_0)$:

$$\Gamma = \frac{1}{n} \sum_{i=1}^{n} 2\mathbf{A}(\mathbf{L}_0)(X_i - X_0)(X_i - X_0)' \mathbf{A}(\mathbf{L}_0^{\perp})' =$$
$$= 2\mathbf{A}(\mathbf{L}_0) \Sigma_{\mathbf{X}} \mathbf{A}(\mathbf{L}_0^{\perp})',$$

где $\Sigma_{\mathbf{X}}$ — ковариационная матрица выборки $\mathbf{X}^{(n)}$. Таким образом, условие экстремальности подпространства $\mathbf{L}_0$ записывается в виде

$$\mathbf{A}(\mathbf{L}_0) \Sigma_{\mathbf{X}} \mathbf{A}(\mathbf{L}_0^{\perp})' = 0 \qquad (20.37)$$

или в терминах базиса Θ_0:

$$\theta_i' \Sigma_{\mathbf{X}} \theta_{i+j} = 0$$

для всех $1 \leqslant i \leqslant q$, $1 \leqslant j \leqslant p - q$.

Рассматривая $(p \times p)$-матрицу $\Sigma_{\mathbf{X}}$ как преобразование пространства R^p, получаем, что условие (20.37) выполняется тогда и только тогда, когда $\Sigma_{\mathbf{X}}$ переводит подпространство

$\mathbf{L}_0$ в себя, т. е. когда $\mathbf{L}_0$ — собственное подпространство ковариационной матрицы[1].

Как показано выше, ортогональная проекция $\mathbf{A}$ ($\mathbf{L}_0$) из R^p в R^q является выразительной относительно F ($\mathbf{A}$) тогда и только тогда, когда $\mathbf{L}_0 = \mathbf{L}$ ($\mathbf{A}$) является экстремальной точкой функции Φ ($\mathbf{L}$).

Таким образом, для проекционного индекса F ($\mathbf{A}$), построенного по статистике $\frac{1}{n}\sum_1^n ||Y_t - Y_0||^2$ на q-мерных выборках, выразительными являются ортогональные проекции на собственные q-мерные подпространства ковариационной матрицы $\mathbf{\Sigma_X}$ и только они.

Для $q = 1$ этот результат лежит в основе метода главных компонент (см. гл. 13). Нетрудно показать, что подмножество $\mathbf{L}_0$ является собственным для матрицы $\mathbf{\Sigma_X}$ тогда и только тогда, когда оно натянуто на какие-либо q главных компонент этой матрицы.

Если выборка $\mathbf{X}$ разбита на k классов (подвыборок) $\mathbf{X}_l = \mathbf{X}$ $(n_l) = (X_{1l}, \ldots, X_{n_l, l})$; $l = 1, \ldots, k$, то определены:

$\mathbf{W_X} = \sum_{l=1}^{k} \pi_l \mathbf{\Sigma_{X_l}}$ — матрица внутриклассового рассеивания;

$\mathbf{B_X} = \sum_{l=1}^{k} \pi_l (X_{0l} - X_0)(X_{0l} - X_0)'$ — матрица межклассового рассеивания. Отметим, что $\mathbf{W_X} + \mathbf{B_X} = \mathbf{\Sigma_X}$. Соответственно определены: внутриклассовый разброс: $S_{\text{в}\mathbf{X}} = \text{Sp}\ \mathbf{W_X}$; межклассовый разброс: $S_{\text{м}\mathbf{X}} = \text{Sp}\,\mathbf{B_X}$, $S_{\text{в}\mathbf{X}} + S_{\text{м}\mathbf{X}} = S_{\mathbf{X}}$.

Пример 20.6. Пусть $\mathbf{X} = \bigcup_{l=1}^{k} \mathbf{X}_l$. Рассмотрим проекционный индекс F ($\mathbf{A}$) для статистики

$$\varphi\left(\mathbf{Y} = \bigcup_{l=1}^{k} \mathbf{Y}_l\right) = \frac{S_{\text{в}\mathbf{Y}}}{S_{\text{в}\mathbf{Y}} + S_{\text{м}\mathbf{Y}}},$$

где $\mathbf{A}$ — проекция из R^p в R^q и $\mathbf{Y}_l = \mathbf{A}\mathbf{X}_l$. Ясно, что F ($\mathbf{A}$) является для всех $q \geqslant 1$ O-инвариантным, а для $q = 1$ даже GL-инвариантным, но F ($\mathbf{A}$) не является GL-инвариантным, если $q > 1$. Следовательно, F ($\mathbf{A}$) можно использовать для поиска выразительных одномерных проекций и выразительных q-мерных ортогональных проекций из R^p в R^q. В каждом из случаев возникает оптимизационная задача

[1] В терминах базиса Θ_0 $(p \times q)$-матрица $\mathbf{A}$ ($\mathbf{L}_0$) называется собственной для матрицы $\mathbf{\Sigma_X}$, если существует $(q \times q)$-матрица $\mathbf{C}$, такая, что $\mathbf{A}$ ($\mathbf{L}_0$) $\mathbf{\Sigma_X} = \mathbf{C}\,\mathbf{A}$ ($\mathbf{L}_0$) При $q = 1$ это соответствует обычному определению собственного вектора.

на многообразии Грассмана для функции, которая в точке $\mathbf{L} = \mathbf{L}(\Theta)$, где $\Theta = (\theta_1, \ldots, \theta_p)$ — ортонормированный базис, вычисляется по формуле

$$\Phi(\mathbf{L}) = \frac{\Phi_{B\chi}(\mathbf{L})}{\Phi_{\chi}(\mathbf{L})} = \frac{\operatorname{Sp} \mathbf{A} W_{\chi} \mathbf{A}'}{\operatorname{Sp} \mathbf{A} \Sigma_{\chi} \mathbf{A}'}, \tag{20.38}$$

где $\mathbf{A} = \mathbf{A}(\mathbf{L})$ — матрица проекции, составленная из q векторов строк $\theta_1, \ldots, \theta_q$.

Имеем:

$$\operatorname{grad} \Phi(\mathbf{L}) = \frac{\Phi_{\chi}(\mathbf{L}) \operatorname{grad} \Phi_{B\chi}(\mathbf{L}) - \Phi_{B\chi} \operatorname{grad} \Phi_{\chi}(\mathbf{L})}{\Phi_{\chi}(\mathbf{L})^2}.$$

Следствие 20.4. Для проекционного индекса $F(\mathbf{A})$, соответствующего отношению усредненного внутриклассового разброса к общему разбросу, наиболее выразительные ортогональные проекции задаются матрицами, составленными из собственных векторов симметрической матрицы $W_{\chi} - \gamma \Sigma_{\chi}$, где $\gamma = F(\mathbf{A})$.

Описанная выше итерационная процедура оптимизации на многообразии Грассмана $G(p, q)$, примененная к функции

$$\Phi(\mathbf{A}) = \frac{\operatorname{Sp}(\mathbf{A} W_{\chi} \mathbf{A}')}{\operatorname{Sp}(\mathbf{A} \Sigma_{\chi} \mathbf{A}')},$$

позволяет отыскать такие выразительные проекции.

Рассмотрим методы оптимизации проекционных индексов $F(\mathbf{A})$ на множестве всех проекций $M(p, q)$ из R^p в R^q.

Пусть $\Pi(q)$ обозначает совокупность всех положительно определенных симметричных $(q \times q)$-матриц. Ставя в соответствие паре матриц $(\mathbf{M}, \mathbf{A})$, где $\mathbf{M} \in \Pi(q)$, а $\mathbf{A} \in MO(p, q)$, матрицу $\mathbf{B} = \mathbf{MA}$, получаем взаимнооднозначное соответствие между множеством всех таких пар

$$\Pi(q) \times MO(p, q) = \{(\mathbf{M}, \mathbf{A}),\ \mathbf{M} \in \Pi(q),\ \mathbf{A} \in MO(p, q)\}$$

и множеством всех проекций $M(p, q)$ из R^p в R^q. Обратное отображение из $M(p, q)$ в $\Pi(q) \times MO(p, q)$ ставит в соответствие проекции $\mathbf{B}$ пару матриц $\mathbf{M} = (\mathbf{BB}')^{1/2}$ и $\mathbf{A} = (\mathbf{BB}')^{-1/2}\mathbf{B}$. Таким образом, каждому проекционному индексу $F(\mathbf{B})$ соответствует функция $F(\mathbf{M}, \mathbf{A})$, и задача поиска выразительной проекции $\mathbf{B}^*$ сводится к оптимизации этой функции на $\Pi(q) \times MO(p, q)$.

Пусть $\mathbf{C}$ — некоторая ортогональная $(q \times q)$-матрица. Тогда если проекции $\mathbf{B}$ соответствует пара $(\mathbf{M}, \mathbf{A})$, то, как легко видеть, проекции $\mathbf{CB}$ соответствует пара $(\mathbf{CMC}', \mathbf{CA})$.

Следовательно, задача поиска выразительной проекции для O-инвариантного проекционного индекса $F(\mathbf{B})$ сводится к задаче оптимизации функции $F(\mathbf{M}, \mathbf{A})$ на следующем множестве классов эквивалентности пар $(\mathbf{M}, \mathbf{A})$:

$$\Pi G(p, q) = \{(\mathbf{M}, \mathbf{A}) : \mathbf{M} \in \Pi(q),\ \mathbf{A} \in MO(p, q) :$$

$$: (\mathbf{M}_1, \mathbf{A}_1) \sim (\mathbf{M}_2, \mathbf{A}_2),$$

тогда и только тогда, когда $\mathbf{M}_1 = \mathbf{C}\mathbf{M}_2\mathbf{C}'$, $\mathbf{A}_1 = \mathbf{C}\mathbf{A}_2$ для некоторой ортогональной матрицы $\mathbf{C}\}$.

Считая координатами симметрической матрицы $\mathbf{M}$ ее матричные коэффициенты $m_{s,t}$, $1 \leqslant s \leqslant t \leqslant q$, можно отождествить множество $\Pi(q)$ с соответствующим подмножеством в $\frac{q(q+1)}{2}$-мерном евклидовом пространстве $R^{\frac{q(q+1)}{2}}$. Условие положительной определенности матриц показывает, что это подмножество является открытой выпуклой областью в $R^{\frac{q(q+1)}{2}}$. Следовательно, для вычисления градиента функции $F(\mathbf{MA})$ по $\mathbf{M}$ можно использовать обычные правила дифференцирования по матричным коэффициентам $m_{s,t}$ матрицы $\mathbf{M}$. Вычисление градиента $F(\mathbf{MA})$ по $\mathbf{A}$ проводится по описанному выше правилу (см. формулу (20.33)).

ВЫВОДЫ

1. Описаны теоретические результаты, лежащие в основе анализа p-мерных случайных величин в терминах их q-мерных проекций, $1 \leqslant q < p$.

2. Исследован класс радиальных распределений. Каждое из этих распределений полностью восстанавливается по единственной одномерной проекции, поэтому смеси их дают запас многомерных модельных законов, достаточный для решения большинства практических задач восстановления плотности распределения по конечному набору проекций.

3. Выделены два важных параметрических семейства радиальных законов $R_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p)$, $\alpha > 0$ и $t_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p)$, $\alpha > p + 1$, где $X \in R^p$ и σ^2 — дисперсия. Каждая q-мерная ортогональная проекция из R^p в R^q переводит $R_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p)$ в $R_{q,\ \alpha+\frac{p-q}{2}}(X; 0, \sigma^2\mathbf{I}_q)$, а $t_{p,\alpha}(X; 0, \sigma^2\mathbf{I}_p)$ в $t_{q,\alpha-(p-q)}(X; 0, \sigma^2\mathbf{I}_q)$, поэтому, задав модель восстанавливаемого закона распределения в виде смесей этих законов, получаем, что каждая его проекция имеет ту же модель со

сдвигом параметров. Большую роль играет и то, что имеется естественный механизм формирования p-мерных векторов с плотностями $R_{p,\alpha}$ и $t_{q,\alpha}$.

4. Для каждого $\alpha > 0$ построен функционал $h_\alpha^*(f_\xi)$ на множестве всех плотностей p-мерных случайных векторов ξ с невырожденной ковариационной матрицей, достигающий максимальное значение на плотности $R_{p,\alpha}(X;\ 0,\ \sigma^2 1_p)$. При $\alpha \to \infty$ функционал $h_\alpha^*(f_\xi)$ переходит в классический энтропийный функционал.

5. Показано, как поиск выразительных проекций из R^p в R^q для данного закона распределения (данной выборки) сводится к решению оптимизационных задач на соответствующих подмногообразиях многообразия всех проекций из R^p и R^q, и описаны алгоритмы решения этих задач.

Глава 21. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ЗАДАЧ СОКРАЩЕНИЯ РАЗМЕРНОСТИ И КЛАССИФИКАЦИИ

Задачи сокращения размерности и классификации часто возникают при обработке данных в различных областях науки и техники. К настоящему времени разработано и продолжает развиваться обширное программное обеспечение (ПО) для решения этих задач.

Подробный обзор программных средств для решения подобных задач, а также для обеспечения других разделов статистического анализа приведен в [143], где рассмотрено значительное число программных продуктов, разработанных у нас в стране и за рубежом. Сведения о программных средствах, полезных в данном разделе статистического анализа, имеются в [12]. Почти все описанные там пакеты и системы статистической обработки данных имеют в своем составе процедуры для сокращения размерностей и классификации. (См. также [66, 75, 89, 95, 120, 203, 204, 249, 256].)

В настоящей главе основное внимание уделено программному обеспечению для персональных ЭВМ (при этом рассматривается программное обеспечение не только для задач сокращения размерностей и классификации, но и для других разделов статистического анализа), а также развитию экспертных систем в статистике.

Рассматриваются также программные средства, предназначенные для таких сравнительно новых подходов в статистическом анализе, как визуализация многомерных данных, разведочный анализ.

21.1. Программное обеспечение прикладного статистического анализа для ПЭВМ

Рассмотрим статистическое ПО в основном для ПЭВМ типа IBM PC и совместимых с ними. В настоящее время статистическое ПО ПЭВМ является весьма развитым. Обзор его по состоянию на 1985 г. приведен в [309]. Здесь же ограничимся рассмотрением сравнительно небольшого списка программных средств, которые, судя по литературным источникам [255] и некоторому нашему личному опыту, представляются наиболее интересными. Данные по ПО сосредоточены в нескольких таблицах, которые представляют характеристики ПО по следующим позициям

Использование ресурсов. В табл. 21.1 представлены характеристики рассматриваемых пакетов. Следует учесть, что разные версии одного и того же пакета могут так же сильно отличаться друг от друга, как и разные пакеты. В графе RAM (random access memory) приведен минимальный объем внутренней памяти, необходимый для работы пакета. В графе «Твердый диск» приводятся две цифры: первая — ми-

Таблица 21.1

Пакет	Версия	RAM (Кб)	Твердый диск	Сопроцессор	Максимальное число	
					переменных	объектов
ABC	2.2	256	—1	0	200	D
BMDP/PC (basis)	PC	640	+1—3	+	300	*D*
BMDP/PC (full)	PC	640	+1—20	+	300	*D*
NCSS	4.2	196	—0.7	0	250	32K
PC-ISP	—	640	0.7	0	200	250K
P-STAT	8 21	640	+2—4	+	150	D
SPSS/PC+(basis)	PC+	384	+1—3	0	200	*D*
SPSS/PC+(full)	PC+	450	+2—6	0	200	*D*
STATA (basis)	1.3	256	—0 3	0	RAM	250K
STATA (full)	1.3	256	—0.4	0	RAM	250K
STATGRAPHICS	1.2, 2.0	384	+1	0	RAM	650K
SYSTAT	2.2	256	+1—2	0	200	D

иимальная память на диске (Мб), необходимая для работы пакета, и вторая — максимальная, запрашиваемая для работы только некоторых программ. Знак «+» означает необходимость диска, «·» — желательность его, «—» — ненужность.

В графе «Сопроцессор» знак «+» указывает на необходимость сопроцессора Intel 8087 для работы пакета, «0» — его использование носит опциональный характер. Заметим, что использование сопроцессора повышает скорость обработки в среднем в 3 раза.

В графе «Максимальное число объектов» буква *D* означает, что объектов может быть столько, сколько их размещается на диске, число килобайт (К) указывает, что объектов может быть столько, сколько поместится в области памяти такого объема (при заданном числе переменных).

Здесь рассматриваются две версии пакета BMDP — базисная (basis) и полная (full), две версии пакета SRSS/PC + (basis и full) и две версии пакета STATA (basis и full).

Базисная версия BMDP содержит 6 программ, а полная — 28 программ. Базисная версия SPSS/PC + не включает некоторые программы по многомерному анализу данных и имеет существенно сокращенные возможности графического анализа данных.

Базисная версия STATA, в отличие от полной, не содержит графических средств анализа.

Управление пакетом и данными. Некоторые сведения, связанные с этими характеристиками, приведены в табл. 21.2. В графе «Способ управления» указано, каким образом осуществляется управление пакетом — с помощью системы меню или команд. С одной стороны, использование меню проще для пользователя-неспециалиста, с другой — развитая система команд позволяет создавать подготовленному пользователю сложные схемы обработки. В графе «Импорт/экспорт» пакеты оцениваются по их возможности взаимодействовать по данным с другими широко используемыми ПО для ПЭВМ — интегрированными пакетами, «spreed sheet» (типа Lotus 1—2—3), базами данных (dBase II/III и т. д.). Здесь, как и в других графах этой таблицы, «+», означает хорошо развитый и легко доступный для пользователя обмен, «·» — удовлетворительный уровень, «—» — возможность имеется, но реализация достаточно трудна. В графе «Манипуляция» приведены оценки возможностей пакетов по работе с файлами — слияние и разделение файлов по переменным и объектам. В четвертой графе в аналогичной шкале оцениваются возможности, предоставляемые

Таблица 21.2

Пакет	Способ управления	Импорт/экспорт	Манипуляция	Преобразование переменных	Пропуски взвешивание	Документация
ABC	Меню	·	—	·	+	—
BMDP/PC (basis)	Команда	+	—	+	·	·
BMDP/PC (full)	Команда	+	—	+	·	·
NCSS	Меню	·	·	·	·	·
PC-ISP	Команда	+	·	+	·	+
P-STAT	Команда	+	+	+	+	·
SPSS/PC + (basis)	Команда	+	+	·	+	+
SPSS/PC + (full)	Команда	+	+	·	+	+
STATA (basis)	Команда	·	+	+	+	·
STATA (full)	Команда	·	+	+	+	·
STATGRAPHICS	Меню	·	·	·	·	·
SYSTAT	Команда	+	+	+	·	+

пользователю для создания новых переменных, преобразования переменных, перекодировки данных и т.д. В графе «Пропуски» оцениваются возможности по работе с пропусками в данных, присвоения весов объектам. В последней графе оценивается уровень документированности и консультаций (help).

Возможности статистической обработки. Данные о реализации в пакетах процедур статистической обработки приведены в табл. 21.3, 21.4 Знак «—» в этих таблицах указывает на отсутствие соответствующей процедуры, «+» — наличие ее, а знак «·» — на то, что имеются ограниченные возможности. Так, для пакета STATGRAPHICS имеется лишь одна кластер-процедура (метод k-средних) и та реализована для небольшого числа объектов ($n \sim 100$).

Следует отметить, что все пакеты хорошо приспособлены для получения дескриптивной статистики одномерных данных и коэффициентов ассоциации (корреляций разных типов, корреляционных отношений, χ^2-статистик и т.д.). Наименьшими возможностями в этом отношении обладает пакет ABC.

Таблица 21.3

Пакет	Регрессионный анализ	Факторный анализ и главные компоненты	Кластер-анализ	Дискриминантный анализ	Логлинейная модель
ABC	·	—	—	—	—
BMDP/PC (basis)	+	—	—	—	·
BMDP/PC (full)	+	+	+	+	+
NCSS	+	·	—	·	—
PC-ISP	+	+	·	·	—
P-STAT	+	·	—	+	—
SPSS/PC (basis)	+	—	—	—	—
SPSS/PC (full)	+	+	+	+	—
STATA (basis)	·	—	—	—	—
STATA (full)	·	—	—	—	—
STATGRAPHICS	+	·	·	·	—
SYSTAT	+	·	+	·	·

Таблица 21.4

Пакет	Лаговые переменные	Бокса—Дженкинса модель	Сглаживание, тренд	Спектральный анализ
ABC	—	—	—	—
BMDP/PC (basis)	·	—	—	—
BMDP/PC (full)	·	+	+	+
NCSS	·	—	+	—
PC-ISP	·	·	+	+
P-STAT	·	—	—	—
SPSS/PC (basis)	·	—	—	—
SPSS/PS (full)	·	—	—	—
STATA (basis)	·	—	—	—
STATA (full)	·	—	·	—
STATGRAPHICS	·	+	+	+
SYSTAT	·	+	+	+

Наиболее полный набор статистических процедур предоставляет пакет BMDP/PC (full). Самым гибким в отношении управления данными является пакет P-STAT.

STATGRAPHICS — сравнительно медленно работающий пакет, и его лучше использовать на более мощных ЭВМ типа IBM AT.

21.2. Проблемы и опыт создания интеллектуализированного программного обеспечения по многомерному статистическому анализу

21.2.1. Что такое «интеллектуализация программного обеспечения» и почему она нужна в прикладной статистике. Как известно[1], конечной целью общей программы разработки ЭВМ *пятого поколения* является создание компьютеров, в которых будет реализован такой резкий скачок их интеллектуальных возможностей, в результате чего машина сможет непосредственно «понимать» задачу, поставленную перед ней непрофессиональным пользователем на естественном языке, т. е. с помощью речи, чертежей, схем, графиков и т.п.

В этой общей программе можно выделить четыре основных направления разработок:

1) развитие *элементной базы* (в частности, уже сегодня реально решение задачи достижения плотности «упаковки» порядка нескольких тысяч вентилей на одном кристалле);

2) разработка *новой архитектуры* (и в первую очередь архитектуры с многими параллельными потоками команд и обрабатываемых данных, предусматривающей, в частности, использование спецпроцессоров);

3) совершенствование *программной технологии* (и в частности, разработка языков высокого уровня для параллельной обработки данных);

4) *интеллектуализация*, т. е. оснащение ЭВМ системой решения задач и логического мышления, обеспечивающей способность машины к самообучению, ассоциативной обработке информации и получению логических выводов, что в конечном счете позволит резко повысить уровень «дружелюбия» машины по отношению к пользователю.

Именно в русле ключевых задач пятого направления лежат проблемы разного уровня интеллектуализации при-

---

[1] *Симонс Дж.* ЭВМ пятого поколения: компьютеры 90-х годов: Пер с англ. — М.: Финансы и статистика, 1985. — 172 с.

кладного (проблемно- и методо-ориентированного) программного обеспечения (ППО). *Экспертные системы* принято относить к одной из основных форм высшего уровня интеллектуализации ППО. Их создание связано в первую очередь с разработкой методов и средств формализации и ввода знаний в компьютерные системы (круг этих вопросов составляет содержание специальной дисциплины — так называемой «инженерии знаний») и манипулирования введенными знаниями.

Таким образом, проблематику, связанную с разработкой экспертных систем, можно отнести к кругу ключевых вопросов решения общей программы создания ЭВМ пятого поколения. Однако следует подчеркнуть разницу в уровне дружелюбия, характеризующем экспертную систему и ЭВМ пятого поколения: услугами последней смогут пользоваться лица, не имеющие опыта работы с ЭВМ, в то время как для работы с экспертной системой все-таки должна быть определенная профессиональная подготовка.

В дополнение к сказанному необходимо остановиться на еще одном факторе, стимулирующем развитие работ в области создания именно *статистических экспертных систем* (СЭС).

Дело в том, что бурно возрастающие объемы информации, требующие грамотной статистической обработки, и почти столь же интенсивно растущее количество промышленного (и коммерчески распространяемого) статистического программного обеспечения (СПО), в основном в виде специализированных пакетов и библиотек (см., например, [309]), находятся в явном дисбалансе с относительно медленно растущей численностью квалифицированных специалистов в области прикладной статистики. Это общая тенденция, но в СССР она проявляется особенно остро.

В результате катастрофически нарастающее число лиц, не являющихся специалистами в области статистического анализа данных, использует СПО независимо от того, получили ли они одобрение специалистов по прикладной статистике и нужно ли это для успешного решения стоящих перед ним задач. Это в свою очередь является причиной развития опасного процесса роста доли неквалифицированного, порой безграмотно-спекулятивного использования СПО, что приводит к дискредитации аппарата прикладной статистики, наносит вред делу.

Распространение опыта специалистов по прикладной статистике в виде СЭС, нацеленных на подсказки и машинное ассистирование, в первую очередь в области предмодельного (разведочного) анализа данных, выбора подходя-

щих моделей и нужной последовательности применяемых методов, интерпретации промежуточных и конечных результатов статистического анализа [1], позволит в какой-то мере ослабить развитие упомянутого опасного процесса роста неквалифицированного использования СПО и смягчить причину этого процесса-дисбаланса между потребностью в квалифицированных специалистах по прикладной статистике и их фактическим наличием.

И наконец, *о социальном аспекте* проблемы создания СЭС. В этой связи следует упомянуть о наличии (в рядах специалистов по прикладной статистике) определенной доли скептиков и даже явных противников, которые считают, что СЭС снижают потребность в знаниях живых специалистов, в какой-то мере заменяют и вытесняют их, выступают в качестве их конкурентов; следовательно, необходимо устраниться от участия в работах по созданию СЭС.

В действительности СЭС позволяет существенно повысить лишь средний, так сказать «ширпотребовский», уровень использования статистических методов анализа данных. Им в настоящее время обладает выросшая в последние десятилетия целая армия особого рода пользователей — «смежников», которые, как правило, «понемногу» ориентируются и в предметной области, в рамках которой решаются соответствующие статистические задачи (в экономике, социологии, медицине, геологии, технике и т.д.), и в инструментарии прикладной статистики, не являясь профессионалами ни там, ни здесь. Вот для этой армии работников кондиционные СЭС действительно представляют угрозу, так как при наличии хороших СЭС этих работников с пользой для дела целесообразно заменить специалистами-профессионалами соответствующих предметных областей.

Что касается профессионалов-статистиков, то создание и распространение СЭС лишь позволит высвободить часть их рабочего времени, отводимого для выполнения функций специалиста средней квалификации (в основном рутинного характера), и переключить его на решение задач более высокого профессионального уровня. Если к этому добавить продуманную систему экономического стимулирования работ профессионалов-статистиков в области создания СЭС,

---

[1] Все эти вопросы относятся к основным «узким местам» в проведении статистического анализа слабо подготовленным (в области прикладной статистики) пользователем, а стандартные СПО, предоставляя пользователю в первую очередь *набор так называемых счетных модулей*, практически никак не помогают ему в преодолении этих узких мест.

то их заинтересованность в развитии этих работ станет не только профессионально-органичной, но и активной

21.2.2. Интеллектуальные возможности статистической экспертной системы и основные вопросы, возникающие при ее создании. Создатели большинства известных к настоящему времени статистических экспертных систем[1] ставили перед собой задачу обеспечить пользователю СЭС машинное ассистирование по следующему кругу вопросов:

1) подсказки по существующим литературным, методическим и программным материалам, относящимся к специфике решаемой задачи,

2) советы в выработке адекватных исходных допущений о природе обрабатываемых данных и в выборе общего вида модели;

3) предложение «меню» подходящих методов статистической обработки с пояснением (в случае запроса пользователя) их сущности, особенностей, сфер применимости;

4) подсказки в построении технологической цепочки статистических процедур и алгоритмов, из которых должна состоять основная обрабатывающая (счетная) программа, и ее автоматическая реализация на ЭВМ;

5) помощь в проведении осмысления и интерпретации промежуточных и конечных результатов статистического анализа и (в случае необходимости) в выработке корректирующих управляющих команд к проведению дальнейшего статистического анализа;

6) помощь в выборе форм представления результатов проведенного статистического анализа

Основной круг пользователей, на который рассчитаны подобные СЭС, это прикладные статистики и математики разного уровня квалификации, а также специалисты предметных областей (экономисты, социологи, медики, инженеры и т. д.), обладающие вероятностно-статистической подготовкой в объеме экономического или технического вуза.

В процессе создания СЭС разработчикам приходится последовательно анализировать следующие вопросы (и уточнять их решение):

а) На какого именно пользователя (предметная область, уровень квалификации) ориентирована создаваемая статистическая экспертная система, каковы конечные прикладные

---

[1] *Hahn G. J.* More Intelligent Statistical Software and Statistical Expert Systems: Future Directions (with Comment by P. F. Velleman and J. W. Tukey) // Amer. Stat. — 1985. — Vol. 39, 1. — P. 1—16.

цели разработки и требования к уровню ее интеллектуализации?

б) Какова структура функционального наполнения и сценария диалога СЭС?

в) Какова главная концептуальная направленность (базовый методологический принцип) создаваемого машинного ассистирования (консультации в выборе и реализации используемых статистических методов, помощь в выборе стратегии статистического исследования и т. д.)?

г) Какие именно технические средства целесообразно привлечь для реализации создаваемой СЭС?

д) Какие типовые и оригинальные программные средства и алгоритмические языки необходимы для создания СЭС?

е) Какие средства интеллектуального ассистирования и интерактивного режима необходимы для построения СЭС?

ж) В какой мере возможно использование существующих, а в какой — необходима разработка новых методов и средств формализации и ввода знаний в компьютерные системы, манипулирования введенными знаниями?

з) Как проводить апостериорную оценку уровня интеллектуализации созданной СЭС?

21.2.3. Серия методо-ориентированных статистических экспертных систем (серия МОСЭС)[1]. Серия методо-ориентированных статистических экспертных систем состоит из определенного числа автономных СЭС, каждая из которых может быть использована для решения задач различных предметных областей (экономики, социологии, медицины, техники и т. п.), *объединяемых лишь общностью необходимого для их решения статистического инструментария*. Другими словами, каждая отдельная экспертная система серии реализует статистический инструментарий одного из разделов прикладной статистики: СЭС по регрессионному анализу, СЭС по классификации объектов и признаков, СЭС по разведочному статистическому анализу и т. п., и в этом смысле может быть отнесена к *методо-ориентированным*. Допускается включение в серию и отдельных *проблемно- и методо-ориентированных* СЭС, т. е. СЭС, предназначенных для решения задач определенной предметной области. Но при этом они требуют использования лишь однородного статистического инструментария (например, в экономике это могут быть СЭС по решению систем одновременных эконометрических урав-

[1] Серия МОСЭС разработана, развивается и сопровождается в Центральном экономико-математическом институте АН СССР и совместном советско-американском предприятии «Диалог».

иеиий или по построению и анализу производствеиных функций: обе эти системы основаны, в инструментальном плане, на статистическом аппарате регрессионного анализа и анализа временных рядов).

Общность различных автономных СЭС, составляющих серию, заключается в их совместимости, а также в возможности расширяемости серии.

Совместимость различиых компонентов серии состоит в одинаковой ориеитации на тип пользователя и уровень интеллектуализации; общиости базового методологического принципа создаваемого в СЭС машинного ассистирования; общности технических и программно-инструментальных средств, на базе которых создается СЭС; возможности взаимных ссылок (т. е., например, пользователь СЭС по регрессионному анализу в процессе дналога с машнной может получить от нее на какой-то стадии решення своей задачи совет произвести такую-то процедуру статистической обработки с помощью, скажем, СЭС по класснфнкацни из данной серии).

Возможность расширяемости серии, т. е. ее пополнеиия новымн СЭС, сопряжена лишь с необходимостью соблюдения при конструироваинн новой СЭС вышеупомянутых условнй совместимостн.

Нже приводится краткое описание функцнонального наполнення компонентов серии методо-орнентированных экспертных систем — «Серии МОСЭС». При выборе разделов прикладного статистнческого анализа разработчикн руководствовались, помимо профессноальных пристрастнй и имеющихся научных заделов, интересами экономических и социально-экономических приложений.

1. **МОСЭС-АВР** — методо-ориентнрованная статнстическая экспертная система по анализу временных рядов (см. например, [12, гл 12, 17] и др.). Необходнмость текущего, оператнвного анализа динамики показателей, характеризующнх состояние нлн функционирование системы (экономической, технической н т. п.) — одна из наиболее распространенных черт характера деятельности многомиллионной армни плановых и управленческнх работников на разиых иерархических уровнях экономнкн. Такого же тнпа задачн постояиио возникают и в разнообразной практике исследовательской деятельности. Здесь и задачи сглаживания временных рядов, нх разложення на трендовую, периодическую (сезонную) и случайиую составляющие, их экстраполяцнн (прогноз), улавлнвання момеитов и характера резких структурных сдвигов и т.д. Именно на репенне такнх задач нацелена МОСЭС-АВР.

2. **МОСЭС-РАЗВАД** — методо-ориентированная статистическая экспертная система по разведочному анализу данных. В практике статистических исследований сложилась печальная традиция (ей, правда, можно найти объективное историческое объяснение), в соответствии с которой важнейший, ключевой этап формирования и обоснования исходных рабочих допущений, закладываемых в основание модели генерирования обрабатываемых статистических данных, как правило, игнорировался. Схема подобных исследований строилась примерно так: «будем полагать (или «есть основания считать»), что анализируемая регрессионная зависимость линейна и характеризуется независимыми и нормально распределенными случайными остатками. Тогда...». На самом деле обрабатываемые статистические данные могут быть не только не нормальными и не независимыми, но и не однородными (в регрессионном смысле). Именно мимо таких «натяжек» в исходных допущениях и приходилось проходить исследователям. Интенсивно развиваемый в последние 10—15 лет аппарат разведочного анализа и, в частности, такие его методы, как целенаправленное проецирование многомерных данных, как раз и нацелены на всестороннее предварительное «прощупывание» исходных данных с целью формирования адекватных рабочих предположений об их вероятностной и геометрической природе, о механизме их генерирования. К настоящему времени в мире имеются считанные единицы программных продуктов, реализующих этот аппарат (см., например, [143]), и ни одной (по нашим сведениям) экспертной системы. Сказанное мотивирует выбор разведочного анализа в качестве «начинки» для одного из компонентов «Серии МОСЭС». В МОСЭС-РАЗВАД, в частности, реализованы методология и значительная часть математического инструментария, описанного в разделах III и IV данной книги.

3. **МОСЭС-РЕГРАН** — методо-ориентированная статистическая экспертная система по регрессионному анализу. Статистический аппарат, позволяющий выявлять и описывать зависимость некоторого количественного результирующего показателя от набора объясняющих переменных, составляет содержание регрессионного анализа и относится, бесспорно, к наиболее широко и часто эксплуатируемому в разнообразных приложениях статистическому инструментарию. Особая актуальность интерактивного диалогового режима общения с ЭВМ в процессе использования этого аппарата связана с реализацией таких его слабо формализованных этапов, как подбор подходящих преобразований для переменных модели, выбор ее общего вида, исследование яв-

ления мультиколлинеарности, анализ влияния резко выделяющихся наблюдений и т.п. Именно в эти моменты «беседа» с СЭС и ее подсказки особенно ценны для пользователя. В обоснование мотивировки выбора этого раздела прикладной статистики в качестве «начинки» одного из компонентов «Серии МОСЭС» следует включить и необходимость программно-вычислительной реализации последних теоретико-методических разработок в данной области и весьма высокую частоту ссылок на этот раздел других компонентов «Серии МОСЭС». В МОСЭС-РЕГРАН реализованы методология и математический инструментарий, описанные в [12].

4 **МОСЭС-КЛАСС** — методо-ориентированная статистическая экспертная система по классификации объектов и признаков. Наряду с регрессионным анализом статистические методы классификации (распознавания образов, дискриминантного анализа, автоматической классификации, кластер-анализа и т. п.) относятся к наиболее широко и часто эксплуатируемому в приложениях, и в первую очередь в экономических и социально-экономических приложениях, статистическому инструментарию. Задачи выявления типологии и типообразующих признаков, технической и медицинской диагностики, предварительной обработки массивов информации с целью их разделения на однородные (в определенном смысле) порции и многие др. обслуживаются методами именно этого раздела ПСА. Продвинутость отечественных теоретико-методических разработок в данной области позволяет рассчитывать на достаточно высокую конкурентоспособность (по меньшей мере по своему функциональному наполнению) данного программного продукта. В нем, в частности, реализованы методология и математический инструментарий, описанные в разделах I и II данной книги.

5 **МОСЭС-СЭУ** — проблемно- и методо-ориентированная статистическая экспертная система по решению и анализу систем одновременных эконометрических уравнений. В ней реализованы методология и математический инструментарий, описанные в [12, гл. 14].

6 **МОСЭС-ПАПРОФ** — проблемно- и методо-ориентированная статистическая экспертная система по построению и анализу производственных функций. Производственные функции, как известно, позволяют в сжатой математической форме представить характерные для анализируемой экономической системы (предприятия, отрасли, всего народного хозяйства) соотношения между объемом выпускаемой продукции, с одной стороны, и размерами основных производственных ресурсов (включая факторы научно-техниче-

ского прогресса) — с другой. Используемый для их построения и анализа статистический аппарат — это регрессионный анализ и анализ временных рядов.

Конечный пользователь «Серии МОСЭС» и характер ее интеллектуализации. «Серия МОСЭС» адресуется как статистику, так и пользователю нестатистику, который, с одной стороны, уже располагает постановкой задачи и четко представляет себе конечные прикладные цели исследования, а с другой стороны, может иметь лишь общее поверхностное представление об аппарате прикладной статистики (на уровне знания основных определений и понятий, таких, как модель регрессии и назначение регрессионного анализа, временный ряд и его тренд, содержание задачи классификации в условиях наличия или отсутствия обучающих выборок, многомерное наблюдение и его проекция на плоскость и т. п.). В эту категорию пользователей попадает, в частности, значительная доля (более 50 %) специалистов той предметной области, к которой относится решаемая задача. Таким образом, среди пользователей «Серии МОСЭС» могут быть как статистики (разного уровня квалификации), так и нестатистики — специалисты соответствующих предметных областей (экономисты, социологи, инженеры, медики и т.д.), имеющие минимальную статистическую подготовку.

При пояснении характера и направленности интеллектуализации описываемых МОСЭС примем следующее условное разложение технологии статистического исследования на элементы.

Элемент 1 (*стратегически-постановочный*): уточнение постановки задачи и конечных прикладных целей исследования.

Элемент 2 (*тактико-методический*): выбор подходящего статистического инструментария, включая определение состава и последовательности реализации статистических процедур, используемых для обработки исходных данных.

Элемент 3 (*счетный*): вычислительная реализация выбранного комплекса методов статистического анализа данных.

Элемент 4 (*интерпретационный*): интерпретация промежуточных и итоговых результатов статистической обработки данных, формулировка выводов, в том числе по поводу направлений дальнейших исследований.

Из этих четырех основных элементов технологии статистического исследования экспертные системы «Серии МОСЭС» претендуют на частичную автоматизацию и машин-

ное ассистирование лишь трех последних: тактико-методического, счетного и интерпретационного. При этом *акцент делается на помощь пользователю в выработке адекватных исходных допущений (гипотез) о вероятностной и геометрической природе обрабатываемых статистических данных и в правильном подборе и описании модели, генерирующей эти данные* («МОСЭС-РАЗВАД» целиком предназначена для решения этих вопросов, а в остальных компонентах «Серии» этому аспекту уделяется существенное внимание).

Общая логическая схема построения диалога «пользователь-ЭВМ». Диалог строится в компонентах «Серии» по принципу «от общего к все более узко методо-ориентированному», а именно: на «входе» в систему—«паспорт» задачи;

1-й уровень диалога: *ЭВМ* — «имеет ли задача статистическую природу?»;
пользователь — «да» или «нет»;

2-й уровень диалога: если «нет», работа СЭС заканчивается; если «да», то к какой из нижеперечисленных (в «меню») областей ПСА она относится: регрессионный анализ, классификация, временные ряды и т.д.;

3-й уровень диалога: (при работе, например, с системой «МОСЭС-КЛАСС», т. е. при ответе «классификация» на предыдущем уровне):
«в какой форме представлены исходные данные?» «меню» возможных форм;

4-й уровень диалога: (если данные представлены в виде многомерных наблюдений): «располагаете ли Вы обучающими выборками?»

5-й уровень диалога: если «нет», то «известно ли Вам число искомых классов?»;

6-й уровень диалога: если «нет», то «желаете ли Вы произвести целенаправленное проецирование исходных данных с целью выработки гипотез о возможном числе классов?»,

7-й уровень диалога: если «да», то обратитесь к «МОСЭС-РАЗВАД» и т. д.

Каждый вопрос ЭВМ сопровождается вспомогательным «примечанием — вопросом» типа: «если какое-нибудь из понятий, участвующих в нашем вопросе, требует разъяснения, сделайте соответствующий запрос».

Инструментальные средства, использованные при создании «Серии МОСЭС». Каждый из компонентов «Серии МОСЭС» оперирует с базой знаний, содержащей не более 400–500 правил и утверждений. Это позволило использовать в качестве технической базы персональные компьютеры IBM PC/XT или IBM PC AT (или полностью с ними совместимые 16-разрядные персональные ЭВМ).

В качестве базовых алгоритмических языков использовались языки «С», «LISP» и некоторые другие (специальные). Операционная система — MS-DOS.

ВЫВОДЫ

1. В настоящее время происходит интенсивное развитие ПО статистики для ПЭВМ. Большинство известных зарубежных пакетов программ (среди них BMDP, SPSS, P-STAT, SAS) имеет в настоящее время версии для ПЭВМ (модели IBM PC XT,IBM PC AT, IBM PS-2). Эти пакеты в основном являются приспособлением к возможностям ПЭВМ версий этих пакетов для больших ЭВМ. В частности, по этой причине они в меньшей степени используют возможности ПЭВМ для создания интерактивного взаимодействия с помощью меню и возможности графики. Специально разработанные с учетом возможностей ПЭВМ пакеты, такие, как STATA (full) SYSTAT и STATGRAPHICS, с другой стороны, предоставляют меньшие возможности собственно статистической обработки.

В целом же следует отметить, что ПО по статистике для ПЭВМ сейчас является достаточно развитым, в том числе и для задач сокращения размерностей и классификации.

2. Перспективной линией развития ПО статистики является разработка интеллектуальных систем статистической обработки данных, в том числе статистических экспертных систем, имеющих широкие возможности машинного ассистирования статистиков-исследователей разного уровня подготовки во время проведения статистического исследования. В настоящее время известно несколько образцов экспертных систем в некоторых областях статистического анализа (см. п. 21.2.1). Их можно рассматривать как первые опытные образцы в этом направлении. Развитие интеллектуального ПО для статистики в рамках «Серии МОСЭС», рассмотренной в п. 21.2.2, позволит создать ряд интеллектуальных статистических систем, охватывающих большинство разделов статистической обработки данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Ленин В. И.* Развитие капитализма в России // Полн. соб. соч.— Т. 3. — 791 с
2. *Абусев Р. А.* К задаче классификации групп многомерных нормальных наблюдений // Прикладная статистика· Ученые записки по статистике — М Наука, 1983.— Т. 45.—С 371—372.
3. *Абусев Р А* О сравнении поточечной и групповой классификации в случае многомерного нормального распределения // Статистические методы — Пермь, 1982 — С. 3—7.
4. *Абусев Р. А., Лумельский Я.П* Несмещенные оценки и задачи классификации многомерных нормальных совокупностей // Теория вероятностей и ее применения — 1980.—№ 2. — С. 381—389.
5. *Айвазян С. А.* Многомерный статистический анализ в социально-экономических исследованиях // Экономика и математические методы. — 1977.-- Т 13 — Вып. 5. — С. 968—985.
6. *Айвазян С. А* Об опыте применения экспертно-статистического метода построения неизвестной целевой функции // Многомерный статистический анализ в социально-экономических исследованиях — М. Наука, 1974.— С. 56—86.
7. *Айвазян С. А.* Статистическое исследование зависимостей.— М.: Металлургия. 1968.—227 с
8. *Айвазян С А.* Экстремальная формулировка основных проблем прикладной статистики //Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа: Всесоюз. школа, Ереван, сент. 1979. — С. 24—49.
9. *Айвазян С. А., Бежаева З. И., Староверов О.В.* Классификация многомерных наблюдений. — М.. Статистика. 1974 —240 с.
10. *Айвазян С. А., Бухштабер В. М* Анализ данных, прикладная статистика и построение общей теории автоматической классификации // Методы анализа данных /Пер. с фр. — М.: Финансы и статистика, 1985.— Вступ ст. — С 5—22.
11. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. — М. Финансы и статистика, 1983.— 472 с.
12. *Айвазян С. А., Енюков И С., Мешалкин Л. Д.* Прикладная статистика. Исследование зависимостей.— М.: Финансы и статистика, 1985 — 488 с.
13. *Айзерман М. А., Браверман Э.М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970.— 384 с.
14. *Алгоритмы и программы восстановления зависимостей*/ Под ред. В. Н. Вапника. — М.: Наука, 1984.—816 с.

15. *Альберт А.* Регрессия, псевдорегрессия и рекуррентное оценивание /Пер. с англ.— М.. Наука, 1977.—224 с.
16. *Андерсон Т.* Введение в многомерный статистический анализ/ Пер. с англ. — М. Физматгиз, 1963. — 500 с.
17. *Андерсон Т.* Статистический анализ временных рядов /Пер. с англ. — М.: Мир, 1976. — 755 с.
18. *Андерсон Т.* Статистический анализ временных рядов в экономике /Пер. с англ. — М.: Статистика, 1972. — 755 с.
19. *Андрукович П. Ф.* Некоторые свойства метода главных компонент // Многомерный статистический анализ в социально-экономических исследованиях. — М.: Наука, 1974.—С. 189—228.
20. *Архаров Л. В.* О предельных теоремах для характеристических корней выборочных ковариационных матриц при больших размерностях // Статистические методы классификации. — М.: Изд-во МГУ, 1972.
21. *Афифи А., Эйзен С* Статистический анализ. Подход с использованием ЭВМ /Пер. с англ. — М.: Мир, 1982 — 488 с.
22. *Барабанюк Т., Шлезингер М.* Об одном алгоритме обучения и оценке качества обучения //Статистические проблемы управления. — Вильнюс, 1976. — Вып 14. — С. 93—104.
23. *Барсов Д. М.* Минимизация ошибки классификации при использовании смещенных дискриминантных функций // Статистика, вероятность, экономика.— М.. Наука, 1985.— С 376—379.
24. *Благовещенский Ю. Н.* Классификация упорядоченных классов // Применение многомерного статистического анализа в экономике и оценке качества продукции: Тез. докл. II Всесоюз. науч.-техн. конференции, Тарту, 1981 — С. 255—261.
25. *Благовещенский Ю. Н., Мешалкин Л. Д.* Линейная классификация распределений с поверхностями постоянного уровня плотности, состоящими из концентрических эллипсов // Статистические методы классификации. — М.: Изд-во МГУ, 1969.— Вып. 1. — С 21—24.
26. *Благовещенский Ю. Н., Мешалкин Л. Д.* Некоторые объекты и проблемы анализа статистических данных // Применение многомерного статистического анализа в экономике и оценке качества продукции: Тез. докл. III Всесоюз. науч.-техн. конференции, Тарту, 1985 — Ч. I. — С. 27—32.
27. *Блехер П. М., Кельберт М. Я.* Доказательство сходимости алгоритма «Форель» // Прикладной многомерный статистический анализ. — М.: Наука, 1978. — С. 358—361.
28. *Бонгард М. М.* Проблемы узнавания. — М.: Наука, 1967. — 320 с.
29. *Боннер Р. Е.* Некоторые методы классификации // Автоматический анализ изображений. — М.: Мир, 1969. — С. 205—234.
30. *Бородин А. И., Бугай А. С.* Биографический словарь деятелей в области математики. — Киев: Радянська школа, 1979. — 607 с.
31. *Боярский А. Я.* Курс демографии. — М.: Статистика, 1975.— 454 с.
32. *Браверман Э. М.* Метод потенциальных функций в задаче обучения машин распознаванию образов без учителя // Автоматика и телемеханика. — 1966.— № 10. — С. 100—121.
33. *Браверман Э. М.* Методы экстремальной группировки параметров и задача выделения существенных факторов // Автоматика и телемеханика. — 1970. — № 1. — С. 123—132.
34. *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983.— 464 с.

35. *Бухштабер В. М., Векслер Л. С.* Автоматическая классификация, основанная на принципе решета Эратосфена // Применение многомерного статистического анализа в экономике и оценке качества продукции: Тез. докл. III Всесоюз. науч.-техн. конференции, Тарту, 1985. — Ч. II. — С. 94—97.
36. *Бухштабер В. М., Маслов В. К.* Векторная оптимизация в СОИИ//Измерительная техника — 1977.— № 5. — С. 46—49.
37. *Бухштабер В. М., Маслов В. К.* Задачи прикладной статистики как экстремальные задачи на нестандартных областях //Алгоритмическое и программное обеспечение прикладного статистического анализа Ученые записки по статистике.— М.: Наука, 1980 — Т. 36. — С. 381—395.
38. *Бухштабер В. М., Маслов В. К.* Факторный анализ и экстремальные задачи на многообразиях Грассмана //Математические методы решения экономических задач. — М.: Наука.— 1977 — № 7. — С. 87—102.
39. *Бухштабер В. М., Маслов В. К.* Факторный анализ на многообразиях и проблема выделения признаков в распознавании образов //Изв. АН СССР. — (Техн. кибернетика). — М., 1975.— № 6. — С. 194—201.
40. *Бухштабер В. М., Маслов В. К.* Томографические методы анализа данных // Применение многомерного статистического анализа в экономике и оценке качества продукции: Тез. докл. III Всесоюз. науч.-техн. конференции, Тарту, 1985.—Ч. I. — С. 33—42.
41. *Бухштабер В. М., Маслов В. К., Зеленюк Е. А.* Методы анализа и построение алгоритмов автоматической классификации на основе математических моделей // Прикладная статистика: Ученые записки по статистике.— М.: Наука, 1983.— Т. 45. — С. 126—144.
42. *Бухштабер В. М., Маслов В. К., Зеленюк Е. А.* Методы построения алгоритмов эталонного типа в задачах автоматической классификации // Машинные методы обнаружения закономерностей: Вычислительные системы. — Новосибирск, 1981. — Вып. 88. — С. 65—79.
43. *Вайнцвайг М. Н.* Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов. — М.:Сов. радио, 1973. — С. 8—12.
44. *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 447 с.
45. *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1973.— 416 с.
46. *Ватанабе С.* Разложение Карунена — Лоэва и факторный анализ: Теория и приложения // Автоматический анализ сложных изображений. — М.: Мир, 1969. — С. 163—181.
47. *Волконский В. А.* Экономико-математические модели согласованного планирования платежеспособного спроса и розничных цен //Экономика и математические методы. — 1973. — Т. IX. — № 4. — С. 16—25.
48. *Выханду Л. К.* Об исследовании многопризнаковых биологических систем // Применение математических подходов в биологии. — Л.: Изд-во ЛГУ, 1964. — С. 19—22.
49. *Гаек П., Гавранек Т.* Автоматическое распознавание гипотез. — М.: Наука, 1984. — 278 с.
50. Герасимова И. А. Структура семьи. — М.: Статистика, 1976.— 176 с.

51. *Гирко В. Л.* Случайные матрицы. — Киев: Радянська школа, 1975.—448 с.
52. *Гранберг А. Г.* Целевая функция общественного благосостояния и критерии оптимальности в прикладных народнохозяйственных моделях // Проблемы народнохозяйственного оптимума. — М.: Экономика, 1969. — С. 78—94
53. *Группировка предприятий отрасли методами теории распознавания образов* // Экономика и математические методы. — 1969. — Т. V.— № 3 — С 353—365.
54. *Гусейн-Заде С.* О задачах таксономии в теоретической географии // Прикладная статистика Ученые записки по статистике — М Наука, 1983. — Т. 45. — С. 378—382.
55. *Деев А. Д.* Представление статистик дискриминантного анализа и асимптотическое разложение при размерности пространства, сравнимой с объемом выборок // Доклады АН СССР. — 1970. — 195. — С. 759—762
56. *Дорофеюк А. А.* Алгоритмы автоматической классификации (обзор) //Автоматика и телемеханика. — 1971. — № 12. — С. 78 — 113.
57. *Дорофеюк А. А.* Алгоритмы обучения машин распознаванию образов без учителя, основанные на методе потенциальных функций // Автоматика и телемеханика. — 1966. — № 10.— С. 78—87.
58. *Дубровский С. А.* Прикладной многомерный статистический анализ. — М.: Финансы и статистика, 1982. — 216 с.
59. *Дуда Р., Харт П.* Распознавание образов и анализ сцен/Пер. с англ. — М . Мир. — 512 с.
60. *Дюран Б., Оделл П.* Кластерный анализ. — М.: Статистика, 1977. — 128 с.
61. *Дэйвисон М* Многомерное шкалирование: методы наглядного представления данных/Пер. с англ.— М.: Финансы и статистика, 1987 — 254 с.
62. *Елисеева И И., Рукавишников В. О.* Группировка, корреляция, распознавание образов. Статистические методы классификации и измерения связей — М.: Статистика, 1977. — 143 с.
63. *Елкина В. Н., Загоруйко Н. Г.* Количественные критерии качества таксономии и их использование в принятии решений// Вычислительные системы. — Новосибирск. Наука. — 1969. — Вып. 36. — С. 32—44.
64. *Енюков И. С.* Методы оцифровки неколичественных признаков // Алгоритмическое и программное обеспечение прикладного статистического анализа: Ученые записки по статистике. — М.: Наука, 1980. — Т. 36. — С. 309—316.
65. *Енюков И. С.* Дискриминантный анализ в системе математического обеспечения обработки данных // Статистика. Вероятность. Экономика.— М.: Наука, 1985. — С. 39—58.
66. *Енюков И. С.* Методы, алгоритмы, программы многомерного статистического анализа (пакет ППСА). — (Матем. обеспечение прикладной статистики). — М.: Финансы и статистика, 1986.—232 с.
67. *Енюков И. С.* Решающие правила на основе линейных комбинаций диагностических показателей // Новости медицинской техники. — 1975. — Вып. 3. — С. 16—25.
68. *Енюков И. С., Кулакова Е. П.* Числовые метки для неколичественных признаков в дискриминантном анализе // Прикладной многомерный статистический анализ. — М.: Наука, 1978.— С. 353—357.

69. *Енюков И. С., Нейштадт А. И.* Оценка дискриминантного пространства по неклассифицированной выборке // Прикладной многомерный статистический анализ. — М.: Наука, 1978.— С. 326—333.
70. *Жирмунская Е. А., Маслов В. К.* Анализ структуры ЭЭГ методами распознавания образов // Физиологический журнал СССР. — Л.: Наука. 1974. — Т. IX. — № 4. — С. 484—490.
71. *Жуковская В. М., Мучник И. Б.* Факторный анализ в социально-экономических исследованиях. — М.: Статистика, 1976. — 151 с.
72. *Журавлев Ю. Г., Торговицкий И. Ш.* Оптимальный метод объективной классификации в задачах распознавания образов // Автоматика и телемеханика. — 1965.— № 11. — С. 2062—2063.
73. Журавлев Ю. И., Камилов М. М., Туляганов Ш. Е. Алгоритмы оценок и их применение. — Ташкент: Фан, 1974. — 124 с.
74. *Завалишин Н. В., Мучник И. Б.* Модели зрительного восприятия и алгоритмы анализа изображений.— М.: Наука, 1974.— 204 с.
75. *Загоруйко Н. Г., Елкина В. Н., Лбов Г. С.* Алгоритмы обнаружения эмпирических закономерностей.— Новосибирск: Наука, 1985. — 110 с.
76. *Заде Л. Л.* Размытые множества и их применение в распознавании образов и кластер-анализе // Классификация и кластер / Пер. с англ. — М.: Мир, 1980. — С. 208—247.
77. *Зангвилл У. И.* Нелинейное программирование /Пер. с англ. — М.: Сов. радио, 1973. — 312 с.
78. *Заруцкий В. И.* Классификация нормальных векторов простой структуры зависимостей в пространстве большой размерности // Прикладной многомерный статистический анализ. — М.: Наука, 1978 — С. 37—51.
79. *Заруцкий В.И.* О выделении некоторых графов связей для нормальных векторов в пространстве большой размерности // Алгоритмическое и программное обеспечение прикладного статистического анализа: Ученые записки по статистике. — М.: Наука, 1980. — Т. 36. — С. 189—208.
80. *Иберла К.* Факторный анализ. — М.: Статистика, 1980. — 389 с.
81. *Каменский В. С.* Методы и модели неметрического шкалирования // Автоматика и телемеханика. — 1977. — № 8. — С. 118—156.
82. *Карп В. П., Кунин П. Е.* Метод направленного обучения в переборной схеме М. М. Боигарда и онкологическая диагностика // Моделирование обучения и поведения.— М.: Наука, 1975. — С. 7 — 17.
83. *Классификация и кластер* /Под ред. Дж. Вэи Райзина. — М.: Мир, 1980. — 390 с.
84. *Климов Г. П.* Инвариантные выводы в статистике. — М.: Изд-во МГУ, 1973. — 186 с.
85. *Кокрен У.* Методы выборочного исследования /Пер. с англ. — М.: Статистика, 1976. — 440 с.
86. *Колмогоров А. Н., Фомин С. В.* Элементы теории функций и функционального анализа. — М.: Наука, 1981.—542 с.
87. *Конотопский В. Ю.* Экспертно-статистический метод построения интегрального показателя экономической эффективности деятельности промышленного предприятия: Дис. ... канд. экон. наук. — М.: МГУ им. М. В. Ломоносова, 1986. —171 с.

88. *Крамер Г.* Математические методы статистики. — М.: Мир, 1975. — 648.
89. *Краскэл Дж. Б.* Многомерное шкалирование и другие методы поиска структуры // Статистические методы для ЭВМ/Пер. с англ.; Под ред. М. Б. Малютова. — М.: Наука, 1986. — С. 301—347.
90. *Краскэл Дж. Б.* Взаимосвязь между многомерным шкалированием и кластер-анализом // Классификация и кластер /Пер. с англ — М. Мир, 1980.— С. 21—41.
91. *Кульбак С.* Теория информации и статистика /Пер. с англ. — М.. Наука, 1967.—408 с.
92. *Куперштох В. Л., Миркин Б. Г., Трофимов В. А.* Сумма внутренних связей как критерий качества классификации // Автоматика и телемеханика. — 1976. — № 3. — С. 91—98.
93. *Лбов Г. С.* Выбор эффективной системы зависимостей признаков // Вычислительные системы. — Новосибирск, 1965.— Вып. 19. — С. 21—24.
94. *Лбов Г. С.* Логические функции в задачах эмпирического предсказания // Эмпирическое предсказание и распознавание образов Вычислительные системы. — Новосибирск, 1978.— Вып. 76. — С. 34—64.
95. *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск· Наука, 1981. — 160 с.
96. *Лоули Д., Максвелл А.* Факторный анализ как статистический метод. — М.: Мир, 1967. — 144 с.
97. *Лумельский В. Я.* Агрегирование объектов на основе квадратичной матрицы // Автоматика и телемеханика. — 1970. — № 1.— С 133—143.
98. *Лумельский В. Я.* Агрегирование матрицы межотраслевого баланса с помощью алгоритма диагонализации матрицы связи // Автоматика и телемеханика. — 1970. — № 9. — С. 69—72.
99. *Льюс Р. Д., Райф Х.* Игры и решения. — М.: Изд-во иностранной лит-ры, 1961. — 643 с.
100. *Макаров В. Л., Айвазян С. А., Житков В. А.* Экспериментальная экономика и стенд экономико-математического моделирования // Экономика и математические методы. — 1989. — т. XXIV. — № 6.
101. *Малиновский Л Г.* Классификация объектов средствами дискриминантного анализа. — М. Наука, 1979. — 260 с.
102. *Маркова Е. В., Маслак А. А.* Рандомизация и статистический вывод. — М. Финансы и статистика, 1986. — 208 с.
103. *Марченко В. А., Пастур Л А.* Распределение собственных значений в некоторых ансамблях случайных матриц // Математический сборник. — 1967. — 72 (114). — № 4. — С. 507—536.
104. *Маслов В. К.* Алгоритмические методы поиска информативных признаков при распознавании случайных процессов // Методы представления и аппаратурный анализ случайных процессов и полей: Тез. докл. V Всесоюз. симпозиума. — Л., 1972.— С. 25—29.
105. *Матюха И. Я.* Статистика бюджетов населения. — М.: Статистика, 1967. — 248 с.
106. *Методы анализа данных*: Подход, основанный на методе динамических сгущений /Пер. с фр.; Под ред. и с предисл. С. А. Айвазяна и В. М. Бухштабера. — М.: Финансы и статистика, 1985.— 357 с. (Математико-статистические методы за рубежом).
107. *Мешалкин Л. Д.* Одновременное изучение динамики прогностической силы некоторых факторов в модели Кокса // Приме-

нение статистических методов в производстве и управлении: Тез. Всесоюз. науч.-техн. конференции. — Пермь, 1984.— С. 132—134.

108. *Мешалкин Л. Д.* Локальные методы классификации // Статистические методы классификации. — М.: Изд-во МГУ, 1969. — Вып. 1. — С. 58—78.

109. *Мешалкин Л. Д., Сердобольский В. И.* Ошибки при классификации многомерных распределений //Теория вероятностей и ее применения — 1978. — С 772—781.

110. *Миркин Б. Г.* Анализ качественных признаков и структур. — М. Статистика, 1980. — 320 с.

111. *Миркин Б. Г.* Группировки в социально-экономических исследованиях. — М. Финансы и статистика, 1985. — 224 с.

112. *Миркин Б. Г.* Метод главных кластеров // Автоматика и телемеханика. — 1987. — № 10. — С. 131—142.

113. *Многомерный статистический анализ в социально-экономических исследованиях.* Ученые записки по статистике /Под ред. С. А. Айвазяна и А. А. Френкеля. — М Наука, 1974. — Т. XXVI — 416 с.

114. *Нейман Дж.* Текущие задачи математической статистики //Международный математический конгресс в Амстердаме, 1954 г. (обзорные доклады)/Пер. с англ. и фр Под ред. С. В. Фомина. — М. Физ.-мат. лит-ра, 1961.— С. 229—258.

115. *Об индивидуальной оценке состояния миокарда в начальных периодах коронарного атеросклероза* (методы многомерного математико-статистического анализа) // Д. Ф. Пресняков, С. А. Айвазян, Ю. А. Розенблат, В. И. Орлов. — Кардиология. — 1975 — № 12. — С 68—82.

116. *Оберхеттингер Ф.* Преобразования Фурье распределений и их обращения Таблицы. — М · Наука, 1979: — 247 с.

117. *Окунь Я.* Факторный анализ./Пер. с польск. — М.: Статистика, 1974 — 200 с.

118. *Онищик А. Л.* Грассмана многообразие // Математическая энциклопедия. — М.: Советская энциклопедия, 1977. — Т.1 — С. 1104—1105.

119. *Орлов А. И.* Некоторые вероятностные вопросы теории классификации // Прикладная статистика.— М.· Наука, 1983. — С. 166—179.

120. *Парринг А. М., Тийт Е. М.* Методическое руководство для пользования пакета САИСИ. — Тарту. Изд-во ТГУ, 1986.— 279 с.

121. *Патрик Э.* Основы теории распознавания образов /Пер. с англ. — М.: Сов. радио, 1980. — 408 с.

122. *Перекрест В. Т.* Нелинейный типологический анализ социально-экономической информации. — Л.: Наука, 1983. — 175 с.

123. *Пикялис В. С* Сравнение методов вычисления ожидаемой ошибки классификации /'Автоматика и телемеханика. — 1976.— № 5. — С. 59—63.

124. *Поляк Б Т.* Введение в оптимизацию. — М.: Наука, 1983.— 384 с.

125. *Программное обеспечение статистической классификации при ограниченных выборках*/Под ред. Ш. Раудиса // Статистические проблемы управления. — Вильнюс. — 1982. — № 58. — 101 с

126. *Прогнозирование надежности изделий электронной техники на основе информативных признаков* /П. С. Гимлявый, В. В. Кобаров, С. А. Колосов и др.; Центр. науч.-иссл. ин-т «Электроника».— М., 1979. — 95 с.

127. *Прохоркас Р. П., Жюжнис В. Е., Мисюнене Н. Б.* Применение некоторых классификаторов для прогнозирования отдаленных исходов инфаркта миокарда // Проблемы ишемической болезни сердца. — Вильнюс, 1976. — С. 261—267.
128. *Развитие сельских поселений* /Под ред. Т. И. Заславской. — М.: Статистика, 1976. — 295 с.
129. *Рао С. Р.* Линейные статистические методы и их применения / Пер. с англ. — М Наука, 1968. — 547 с.
130. *Распознавание образов в социальных исследованиях* / Под ред. Н Г. Загоруйко и Т. Н. Заславской. — Новосибирск: Наука, 1968. — 195 с.
131. *Растригин Л. А., Эренштейн Р. Х.* Метод коллективного распознавания. — М. Энергоиздат, 1981.— С. 1—78.
132. *Раудис Ш.* Ограниченность выборки в задачах классификации// Статистические проблемы управления. — Вильнюс. — 1976. — № 8. — С. 6—185.
133. *Раудис Ш.* Ошибки классификации при выборе признаков // Статистические проблемы управления.— Вильнюс.— 1979. — № 38 — С. 9—25.
134. *Раудис Ш., Пикялис В., Юшкявичус К.* Экспериментальное сравнение тринадцати алгоритмов классификации //Статистические проблемы управления. — Вильнюс. — 1975.— № 11.— С. 53—80.
135. *Раудис Ш., Пикялис В.* Табулирование зависимости ожидаемой ошибки классификации линейной дискриминантной функцией от объема обучающей выборки // Статистические проблемы управления. — Вильнюс. — 1975. — № 11. — С. 81—119.
136. *Раушенбах Г. В.* Проблемы измерения близости в задачах анализа данных //Программно-алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. — М.: Наука, 1987. — С 41—54.
137. *Раушенбах Г. В.* Об измерении близости между множествами в задачах кластер-анализа // Статистика. Вероятность. Экономика. — М. Наука, 1985. — С. 388—392.
138. *Розин Б. Б.* Теория распознавания образов в экономических исследованиях. — М. Статистика, 1973 — 224 с.
139. *Севрюк М. Б.* Сходимость алгоритма «Форель» для бесконечного числа объектов //Алгоритмическое и программное обеспечение прикладного статистического анализа: Ученые записки по статистике. — М.. Наука, 1980. — Т 36. — С. 377—381.
140. *Сердобольский В. И.* Влияние отбора компонент случайной величины на классификацию // Изв. высш. учеб. заведений (Математика). — 1983. — № 9. — С. 46—55.
141. *Сердобольский В. И.* Дискриминантный анализ при большом числе переменных // Доклады АН СССР. — 1980. — 254. — №1. — С. 39—44.
142. *Сердобольский В. И.* О минимальной вероятности ошибки в дискриминантном анализе // Доклады АН СССР. — 1983.— 270. — № 5. — С. 1066—1070.
143. *Сильвестров Д. С.* Программное обеспечение прикладной статистики — М. Финансы и статистика, 1988. —240 с.
144. *Система распознавания речи СИРИУС 2.1*/В. Н. Плотников, В. А. Суханов, Ю. Н. Жигульцев, В. А. Кононенков // Тр. Всесоюз. семинара «Автоматическое распознавание слуховых образов АРСО-13». — Новосибирск. 1984. — Ч. 2. — С. 94.

145. *Староверов О. В.* Интервальный прогноз структуры населения//Экономика и математические методы. — 1976.— Т. XII.— № 1. — С. 56—71.
146. *Староверов О. В.* Модели движения населения. — М.: Наука. 1979.—344 с.
147. *Статистический анализ экспертных оценок* /П. Ф. Андруковнч, Г. Н. Веселая, В. П. Козырев, А. Т. Терехин // Многомерный статистический анализ в социально-экономических исследованиях: Ученые записки по статистике. — М.: Наука, 1974. — Т. XXVI. — С. 168—188.
148. *Степанов В. С.* Об ортогонально инвариантных дискриминантных функциях с асимптотически оптимальными смещенными оценками обратной матрицы ковариаций большой размерности: алгоритм, моделирование, применение /МГУ им. М. В. Ломоносова. Фак. вычисл. матем. и киберн.— М., 1987.— С. 87.— Деп. в ВИНИТИ 10.10.87, № 7374—В 87.
149. *Социально-демографическое развитие села.* Региональный анализ/Под ред. Т. И. Заславской, И. Б. Мучника. — М.: Статистика, 1980.—343.
150. *Статистические методы для ЭВМ* / Под ред. К. Энслейна, Э. Рэлстона, Г. С. Уилфа; Пер. с англ. — М.: Наука, 1986. — 460 с.
151. *Терентьев П. В.* Дальнейшее развитие метода корреляционных плеяд//Применение математических методов в биологии.— 1960. — № 1. — С. 44—62.
152. *Терехина А. Ю.* Анализ данных методами многомерного шкалирования. — М.: Наука, 1986. — 168 с.
153. *Типология и классификация в социологических исследованиях* / Под ред. В. Г. Андреенкова и Ю. Н. Толстовой.— М.: Наука, 1982. — 296 с.
154. *Типология потребления*/Под ред. С. А. Айвазяна и Н. М. Римашевской. — М.: Наука, 1978. — 168 с.
155. *Турбович И. Т., Гитис В. Г., Маслов В. Г.* Опознание образов. — М.: Наука, 1971. — 246 с.
156. *Ту Дж., Гонсалес Р.* Принципы распознавания образов.— М.: Мир, 1978. — 416 с.
157. *Уилкс С.* Математическая статистика/Пер. с англ. — М.: Наука, 1967.— 632 с.
158. *Укрупненные и межотраслевые модели народного хозяйства.* — Новосибирск: Наука, 1976. — 216 с.
159. *Флексер Л. А., Айвазян С. А.* Выявление закона распределения хлопковолокна по длине // Известия высших учебных заведений: технология текстильной промышленности. — 1962. — № 1 (26). — С. 33—41.
160. *Фукунага К.* Введение в статистическую теорию распознавания образов/Пер. с англ. — М.: Наука, 1979. — 367 с.
161. *Харман Г.* Современный факторный анализ /Пер. с англ. — М: Статистика, 1972. — 486 с.
162. *Харман Г.* Восстановление изображений по проекциям. Основы реконструктивной томографии/Пер. с англ. — М.: Мир,1983.— 349 с.
163. *Хелгасон С.* Преобразование Радона/Пер. с англ. — М.: Мир, 1983. — 150 с.
164. *Хеттманспергер Т.* Статистические выводы, основанные на рангах. — М.: Финансы и статистика, 1987. — 334 с.
165. *Ченцов Н. Н.* Статистические решающие правила и оптимальные выводы. — М.: Наука, 1972. — 520 с.

166. *Шлезингер М. Н.* О самопроизвольном различении образов// Читающие автоматы. — Киев: Наукова думка, 1965. — С. 38—45.
167. *Шурыгин А. М.* Пути улучшения линейной дискриминации в нормальном случае //Статистика, вероятность, экономика. — М.: Наука, 1985. — С. 379—382.
168. *Юшкявичюс К. З.* Исследование чувствительности кусочно-линейного классификатора по минимуму расстояния к ограниченности объема обучающей выборки // Статистические проблемы управления. — Вильнюс. — 1983. — № 61. — С. 89—109.
169. *Юшкявичюс К. З., Раудис Ш. Ю.* Вычисление ожидаемой ошибки классификации для кусочно-линейного решающего правила // Статистические проблемы управления. — Вильнюс. — 1976. — № 14. — С. 67—84.
170. *Abend K., Harley T. J.* Comments on the mean accuracy of statistical pattern recognition // IEEE Transactions on Information Theory. — 1969. — Vol. 15. — P. 120—121.
171. *Adanson M.* Histoire naturelle du Sénégal. Coquillages. Avec la relation abrégée d'un voyage fait dans ce pays de 1749 à 1753. Bauche, Paris, 1757.
172. *Agrawala.* Learning a probabilistic teacher // IEEE, Inform. theory. — 1970. — Vol. 16. — N 4.
173. *Aivazian S. A.* Probabilistic — Statistical Modelling of the Distributary Relations in Society // Private and Enlarged Consumption. — Ed. by L. Solari, Q. — N Du Pasquier North — Holland. Publishing Company Amsterdam — New York — Oxford. — 1976. — P. 285—247.
174. *Anderberg M.R.* Cluster analysis for application. — N. Y.: Academic Press, 1973.
175. *Anderson Q. A.* Logistic discrimination // Krishnaiah P. R. (ed.) Handbook of statistics. — Vol. 2: Classification, Pattern recognition and Reduction of dimensionality. — North — Holland. — 1982. — P. 169—191.
176. *Anderson* T. W. Asymptotic theory for component analysis // Ann. Math. Statist. — 1963. — Vol. 34. — P. 122—148.
177. *Anderson T. W.* The asymptotic distribution of characteristic roots and vectors// Proc. 2 Berkeley Symp. Math. Statist. and Probab. — Univ. Calif. Press, 1952.— P. 103—130.
178. *Anderson T. W., Bahadur R. R.* Classification into two multivariate normal distribution with different covariance matrices // Ann. Math. Statist. — 1962. — Vol. 33. — P. 422—431.
179. *Anderson S. et. al.* Statistical methods for comparative studies. — Wiley, 1980.— 283 p.
180. *Anderson T. W., Rubin H.* Statistical inference in factor analysis. // Proc. 3 Berkeley Symp. Math. Statist. and Probab. — Univ. Calif. Press, 1956, 5. — P. 11—50.
181. *Andrukowich P. F.* a. o. Abstract painting as a specific — Generale— Language. A Stat. Appr. to the problem // Metron XXIX.— 1971. — N 1—2.
182. *Arabie P.* Random versus rational strategies for initial configurations in nonmetric multidimensional scaling// Psychometrika. — 1975. — Vol. 43. — N 1.
183. *Asimov D.* The grand tour: A tool for multidimensional data // SIAM Journal of Scientific and Statistical. — Computing. — 1985. — 6. — P. 128—143.
184. *Bahadur R. R.* A representation of the Joint Distribution of Responses to n Dichotomous Items // Solomon H. (ed.) Studies in

Analysis and Prediction. — Stanford, California. — 1961. — P. 158—168.
185. *Bartlett M. S.* Factor analysis in psychology as a statistician sees.— Uppsala: Almqvist and Wiksell, 1953. — P. 23—34.
186. *Bekker R. A., Chambers J. M. S.* An interactive envitroment for data analysis and graphics. — Wadsworth, 1974.
187. *Bekker R. A., Chambers J. M.* Extending the S system. — Wadsworth, 1985.
188. *Benard J.* Quelques aspects theoretique des biens collectivs sous tutelle; Le repport sur «Conférence sur la planification et le marche», Libilic — Tchécoslovaqiue, 4 — 8 Mai, 1970.
189. *Ben — Bassat M.* Use of Distance Measure, Information Measures and Error Bounds in Feature Evaluation // Krishnaiah, P. R. (ed.) Handbook of Statistics. — Vol. 2; Classification, Pattern recognition and Reduction of Dimensionality.— North-Holland.— 1982. — P. 773—791.
190. *Bennet R. S.* The intrinsic dimensionality of signal collections // IEEE Trans. Inform. Theory. — 1969. — Vol. 15. — № 5.
191. *Benzecri J. P.* L'Analyse des Données.— Tome 1; La Taxinomie.— Tome 2; L'Analyse des Correspondances. — Paris; Dunod, 1976 (2nd ed). — 632 p.
192. *Bezdek J. C.* Pattern recognition with fuzzy objective function algorithms // Plenum Press. — N-Y, Z. 1980.
193. *Bezdek J. C.* Numerical Taxonomy with fuzzy sets// J. of Math. Biol. — 1974. — N1. — P. 57—71.
194. *Blagoveshensky Yu. N., Meshalkin L. D.* Multidinensional T-normal distribution // I Всемирн. конгресс об-ва матем. стат. и теории вероятностей им. Бернулли: Тез. докл. — М.: Наука, 1986. — Т.1. — С. 79.
195. *Bolshev L. N.* Cluster analysis // Bull. Int. Stat. Inst. — 1969.— N 43. — P. 441—425.
196. *Box Q. E. P., Cox D. R.* An analysis of transformation / J. R. Statist. Soc., B — 1964 — Vol. 26. — P. 211—252.
197. *Boyles.* On the convergence of the EM algorithm // J. R. Statist. Soc., B. — Vol 45. — N 1. — 1983.
198. *Broffit. J.* Nonparametric classification. // Krisnaiah. P. R. (ed.) Handbook of statistics. — Vol. 2: Classification, Pattern recognition and Reduction of dimensionality. — North-Holland. — 1982. — P. 139—168.
199. *Bryant P., Williamson J.* Asymptotic Behaviour of Classification Maximum Likelihood // Biometrika. — 1978. — Vol. 65.
200. *Burt C.* The Factorial Analysis of Qualitative Data // J. Stat. Psychol. — 1950. — Vol. 3. — N 3. — P. 166—185.
201. *Carrol J. D., Chang J. J.* Analysis of Individual Differences in Multidimensional Scaling via an Generalization of Ecart-Young Decomposition // Psychometrika. — 1970. — Vol. 35.— N 5. — P. 283—319.
202. *Celeux Q., Diebolt J.* Reconnaissance de mélange de densité et classification Un algorithme d'apprentissage probabiliste: l'algorithme SEM // Rapports de Recherche de L' INRIA. Centr de Rocquencort. — 1984.
203. *Compstat — 1984.* Proceedings in Computational Statistisc. — Wien: Physica, 1984. — 520 p.
204. *Compstat — 1986.* Proceedings in Computational Statistics. — Heidelberg: Wien: Physica — Verlag, 1986. — 512 p.
205. *Cooper.* Non supervised adaptative signal detection and pattern recognition // Information and Control. — 1964. — Vol. 7.

206. *Cox D. R.* Regression models and life tables // J. R. Statist. Soc., B. 34. — 187 p.
207. *Cox D. R.* Some remarkes on role in statistics of graphical methods // Applied Statistics. — 1978. — Vol. 27. — P. 4—9.
208. *Cressie N.* The power results for tests based on high — order gaps // Biometrika. — 1978. — Vol. 65. — P. 214—218.
209. *Day N. E.* Divisive cluster analysis and test for multivariate normality // Session of the ISI. — London, 1969.
210. *Day N. E.* Estimating the components of a mixture of normal distributions // Biometrika. — 1969. — Vol. 56. — N 3. — P. 463—474.
211. *De Leeuw J., Pruzansky S.* A new computational method to the weighted euclidean distance model // Psychometrika. — 1978.— Vol. 43. — N 4.
212. *Dempster A., Laird G., Rubin J.* Maximum Likelihood from incomplete data via the EM algorithm // J. R. Statist. Soc. — 1977. — B. — Vol. 39.
213. *Devlin S. J., Gnanadesikan R., Kettering J. R.* Robust Estimation of Dispersion Matrices and Principal Components // J. Amer. Stat. Ass. — 1981. — Vol. 76. — P. 354—362.
214. *Di Battista G., Tamassia R.* An Interactive Graphic System for Disigning and Accessing Statistical Data Bases // Compstat-86, Proceedings in Computational Statistics.—Wien: Physica — Ferlag, 1986. — P. 231—236.
215. *Diaconis P., Freedman D.* Asymptotics of graphical projection pursuit // Ann. Statist. — 1984. — 12. — P. 793—815.
216. *Diday E.* Classification automatique sequentielle pour grands tableaux // Rev. Fr. Inf. Rech. Oper. — 9-e année a mars 1975. — B—1 — P. 29—61.
217. *Di Pillo P. J.* Biased discriminant analysis: evaluation of the optimum probability of misclassification // Commun. Statist.— theor. meth. — 1979. — A8 (14). — P. 1447—1457.
218. *Edwards A. W., Cavalli — Sfora L. L.* A method for cluster analysis // Biometrics. — 1968. — 21. — P. 362—375.
219. *Efron B.* Bootstrap methods another at the jackknife // Ann. Statist. — 1979. — 7. — N 1. — P. 1—26.
220. *Efron B.* The efficiency of logistic regression comprared to normal discriminant analysis // J. Amer. Stat. Ass. — 1975. — 70. — P. 892—898.
221. *Everitt B. S.* Unresolved problems in cluster analysis // Biometrics. — 1979. — Vol. 35. — P. 169—181.
222. *Everitt B. S., Hand D. J.* Finit mixture distribution. — Chapman and Hall, 1981.
223. *Fienberg S. E.* Graphical methods in statistics // The Amer. Stat. — 1979. — Vol. 33. — P. 165—178.
224. *Fisher L., Yonh W.* Wan Ness. Admissible clustering procedures// Biometrika. — 1971. — 58. — N 1. — P. 91—104.
225. *Fix E , Hodges J. L. Jr.* Discriminatory analysis, nonparametric discrimination USA School of Medicine. — Texas: Rendolph Field, 1951, 1952.
226. *Friedmon J. H.* A tree-structured approach to nonparametric multiple regression // Lecture Notes in Mathematics 757, Smoothing Technigues for Curve Estimation. — Berlin Springer. — P. 5—22.
227. *Friedman J. H.* Exploratory Projection Pursuit // J. Amer. Stat. Ass. — 1987. — P. 249—256.

228. *Fridman H. P., Rubin J.* On some invariant criterion for grouping data // J. Amer. Stat. Ass. — 1967. — 67. — P. 1159—1178.
229. *Friedman J. H., Stuetzle W.* Projection pursuit regression // J. Amer. Stat. Ass. — 1981. — 76. — P. 817—823.
230. *Friedman J. H., Tukey J. W.* A projection pursuit algorithm for exploratory data analysis // IEEE Trans. Comput. — 1974. — Vol. — 23. — P. 881—889.
231. *Friedman J. H., Stuetzle W., Schroeder A.* Projection pursuit density estimation // J. Amer. Stat. Ass. — 1984. — 79. — P. 599 — 608.
232. *Fukunaga K., Koontzwarren L. G.* Application of the Karhunen—Loeve expansion to feature selection and ordering // IEEE Trans. Comput. — 1970. — C. 19. — N 4. — P. 311—318.
233. *Fukunaga K., Hostetler L. D.* The extimation of the gradient of a density function with application in pattern recognition // IEEE Trans. Inform. Theory. — 1975. — Vol. 21. — P. 32—50.
234. *Galinski R. B., Harabasz J.* Dendrite method for cluster analysis // Communications in Statistics. — 1974. — Vol. 3. — P. 1 — 27.
235. *Girshik M. A.* On the sampling theory of roots of determinantial equations // Ann. Math. Statist. — 1939. — Vol. 10. — P. 203—224.
236. *Girshik M. A.* Principal components // J. Amer. Stat. Ass. — 1936. — Vol. 31. — P. 519—528.
237. *Gittman I., Levine M. D.* An algorithm for unimodal fuzzy sets and application as a clustering technique // IEEE Trans. Comput. — 1970. — Vol. 19. — N 7. — P. 583—593.
238. *Glick N.* Additive estimators for probabilities of correct classification // Pattern Recognition. — 1978. — 1. — N 3. — P. 211—222.
239. *Gordesch J.* Non — Standard Graphical Presentation // Compstat-86, — Proceedings in Computational Statistics. — Wien: Physica — Ferlag, 1986 — P. 237—242.
240. *Greenacre M. J., Browne M. W.* An efficient alternating least squares algorithm to perform multidimensional unfolding // Psychometrika. — 1986. — Vol. 51. — N 2. — P. 241 — 250.
241. *Geman S.* A limit theorem for the norm of randon matrices // Ann. Probab. — 1980. — Vol. 8. — P. 252—261.
242. *Gupta S.* Optimum classification rules for classification into two multivariate normal population //Ann. Math. Statist. — 1965. — 36. — N 4
243. *Haas de J. H.* Changing of mortalite patterns and cardiovascular diseases. De Erven F. — Bohn, 1964.
244. *Holzinger K., Harman H.* Factor analysis. — Univ. Chicago Press, 1941.
245. *Huber P.* Experiences with three — dimensional scatterplots // I. Amer. Stat. Ass. — 1987. — Vol. 82. — P. 448—453.
246. *Huber P. J.* Projection Pursuit. Invited paper. // The Annals of Statistics. — 1985. — Vol. 13. — N 2. — P. 435—475; Discussion. — P. 475—525.
247. *Hothakker H. S.* Revealed preference and utility function // Econometrika. — 1949. — 17. — N 2. — P. 195.
248 *Jambu M.* Classification automatiger pour l'analyse des données. 1 — méthodes et algorithmes — 310.p. 2 — logiciels — 400 p. Dunod., Bordas — Paris, 1978.

249. *Jambu M.* Fortran IV computer program for rapid hierarchical classification of large data sets // Computers Geosciences. — 1981. — Vol. 7. — P. 297—310.
250. *Jardin N., Sibson R.* The structure and construction of taxonomic hierarchies // Math. Biosciences. — 1967. — A. — P. 173—179.
251. *Jones M. C., Sibson R.* What is projection pursuit // J. R. Statist. Soc. — 1987. — sec. A. 150. — Part 1. — P. 1—36.
252. *Jeffers J. N. R.* Two case studies in the application of principal component analysis // Appl. Stat. — 1967.— 16. — N 3.
253. *Kendall M. G.* Discrimination and Classification. Multivariate Analysis // Proc. Intern. Symp. held in Dayton, June, 1965. — P. 165—185.
254. *Kazakos G.* Recursive estimation of prior probabilities using a mixture // IEEE Inform. Theory. — 1977. — Vol. 23.
255. *Keller W. J.* Statistical vio Personal Computers // Compstat-86, Proceedings in Computational Statisticals. — Wien: Physica — Ferlag, 1986. — P. 332—337.
256. *Kent D. P., Young F.W.* User's guide for VISUALS software for hyperdimensional dynamic graphics/SAS-Institute, 1986.
257. *King B. F.* Stepwise clustering procedures // J. Amer. Stat. Ass. — Vol. 62. — P. 86—101.
258. *Kronmal R. A., Peterson A. V.* A variant of the acceptance — rejection method for computer generation of random variables // J. Amer. Stat. Ass. — 1981. — Vol. 76. — N 374. — P. 446 — 452.
259. *Kruskal J. B.* Linear transformation of multivariate data to reveal clustering // Multidimensional Scaling: Theory and Applications in the Behavioral Sciences. — London: Seminar Press, 1972. — Vol. 1. — P. 179—191
260. *Kruscal J. B.* Monotone Regression· Continuity and Differentiability Properties // Psychometrika. — 1971. — Vol. 36. —N 1.— P. 57—63.
261. *Kruskal J. B., Carrol J. D.* Geometric models and badness-of-fit functions // Multivariate Analysis / Ed. Krishnaiah P. R. — 1969. — Vol. 11. — N 4. — Acad. Press.
262. *Lance G. N., Williams W. T.* A general theory of classificatory sorting strategies // Hierarchical systems. Comp., J. — 1967. — P. 373—380.
263. *Lebart L., Morineane A., Warwick K. M.* Multivariate Descriptive Statistical Analysis // Correspondence Analysis and Related Techniques for Large Matrices. — N. Y., Chichester, etc. J. Wiley & Sons. — 1985. — P. 231.
264. *Lachenburch P. A. and Mickey R. M.* Estimation of error in discriminant analysis // Technometrics. — 1968. — 10.— N 1.— P. 1—11.
265. *Lebart L.* The Significancy of Eigenvalues Issued from Correspondence Analysis // Compstat-76, Proc. Comput. Statist. — Wien. Physica — Verlag, 1976. — P. 38—45.
266. *Lefkoviten L. P.* Conditional clustering // Biometrics. — 1980.— Vol. 36. — P. 43—58.
267. *Logan B. F., Shepp L. A.* Optimal reconstruction of a function from its projections // Duke Math. J. — 1975. — 42. — P. 645—659.
268. *Mac Queen J.* Some methods for classification and analysis of multivariate observations // Proc. Fifth Berkeley Symp. Math. Stat. and Probab. — 1967. — 1. — P. 281—297.

269. *Maronna R. A.* Robust M-Estimators of Multivariate Location and Scatter. // Ann. Statist. — 1967. — Vol. 4. — N 1. — P. 51—67.
270. *Mays R.* Interactive maximum reliability cluster analysis // Educational and Psychological Measurement. — 1978. — Vol. 38. — P. 783—785.
271. *Meshalkin L. D., Kagan A. B.* A contribution to the discussion upon the paper «Regression models and life tables» by D. R. Cox// J. R. Statist. Soc. — 1972. — Ser. B. — N 2.
272. *Meshalkin L. D.* Some mathematical methods for the study noncommunicable diseases // Planning health Services Proceedings of the Sixth International Scientific Meeting. August 29—September 3, 1971.—Savremena administracija.—Belgrade, 1973. — Vol. 1. — P. 250—256.
273. *Milligan G. W.* A Monte-Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis // Psychometrika.—1 — 81.— Vol. 46. — 2.— P. 187—197.
274. *Milligan D. W., Cooper M. C.* An examination of procedures for determening the number of clusters in a data set. // Psychometrika. — 1985. — Vol. 50. — N 2. — P. 159—179.
275. *Mirkin B. G.* Additive clusterring and qualitative factor analysis methods // Journal of Classification.—1 — 87. — 4. —N 1.— P. 7—31.
276. *Molenaar I. W.* Computer Graphics and Data Presentation, a First Step Toward a Cognitive and Ergonomic Analysis // Compstat-86, Proceedings in Computation Statistics. — Wien: Physica—Ferlag, 1986. — P. 243—250.
277. *Morris J. N. et. al.* Incidence and prediction of ischesmic heart disease in London Busmen // Lancet. — 1966. — Vol. 12.—P. 553—559.
278. *Morrison D. G.* Measurement problems in cluster analysis // Management Science. — 13. — P. B775—B780.
279. *Morrison D. G.* Multivariate statistical methods. — N. Y.: Mc Grou Hill Book Company, 1967.
280. *Murthy V. K.* Nonparametric estimation on Multivariate densities with applications // Multiv. Anal. Proc. Intern. Symp. held in Dayton Ohio, June 14—19, 1965.
281. *Nady G.* State of Art Pattern Recognition // Proc. IEEE. — 1968. — Vol. 56. — N 5. — P. 836—862.
282. *Okamoto M.* An asymptotic expansion for the distribution of the linear discriminant function // Ann. Math. Statist. — 1963.— Vol. 34. — N 4. — P. 1296—1301. Correction // Ann. Math. Statist. — 1968. — Vol. 39. — P. 1358—1359.
283. *Okamoto M.* Optimality Principal Components Multivariate Analysis // Proc. 3 Int. Symp. Dayton. — 1967.
284. *Okamoto M., Kanazawa M.* Minimization of Eigenvalues of a matrix and optimality of principal components // Ann. Math. Statist. — 1968. — Vol. 39. — N 3.
285. *Öhrvik J.* Structure in noise. Research Report Department of Statistics. — University of Stockholm, 1987. — 25p.
286. *Ohsumi N.* Practical Use of Color Imaging in Automatic Classification // Compstat-86: Proceedings in Computational Statistics.—Heidelberg, Wien: Physica — Verlag, 1986.— P. 489—497.
287. *O'Neill M. E.* Distributionnal Expansions for Cannonical Correlations from Contingency Tables // J. R. Statist Soc. — 1978. — Vol. 40. — Ser. B. — P. 303—312.

288. *Orloci L.* Information theory models for hierarchic and nonhierarchic classifications // Numerical taxonomy/ed. by. Cole A.J.—N-Y.· Acad. Press. — 1969. — P. 148—164.
289. *Orloci L.* Geometric models in ecology. I. The theory and application of ordination methods //J. Ecology. — 1966.— Vol. 54.— N 1.
290. *Parzen E.* On Estimation of Probability Density Function and Mode // Ann. Math. Statist. — 1962. — Vol. 33. — Sept. N 3. — P. 1065—1076.
291. *Quandt R. E., Ramsey J. B.* Estimating mixtures of normal distributions and switching regression // J. Amer. Stat. Ass. — 1978. — Vol. 73. — P. 730—738.
292. *Rao C. R.* Linear Statistical Inferences and its Applications. — N.-Y.: Wiley, 1965.
293. *Rao C. R.*The use and interpretation of principal components analysis in applied research // Sankhya, A. — 1964. — Vol. 26.— N 4. — P. 329—358.
294. *Rao C. R.* Estimation and tests of significance in factor analysis // Psychometrika. — 1955. — 20. — P. 93—111.
295. *Redner L., Walker J.* Mixture densities, maximum likelihood and the EM algorithm // SIAM Review. — 1984. — Vol. 26. — N 2.
296. *Rubin J.* Optimal classification into groups: an approach for solving taxonomy problem // J. Theor. Biol. — 1967. — 15.— P. 103 — 144.
297. *Ruspini E. H.* A new approach to clustering // fnformation and Control. — 1 — 69. — Vol 5. — P. 22—32.
298. *Ruspini E. H.* Numerical methods for fuzzy clustering // Inform. Sciences. — 1970. — Vol. 2. — N3. — P. 319—350.
299. *Saito T.* The problem of the additive Constant and eigenvalues in metric multidimensional scaling // Psychometrika. — 1978.— Vol. 43. — N 2.
300. *Sammon J. W.* A nonlinear mapping for Data Structure Analysis // IEEE Trans. Comput. — 1969. — C — 18. — N 5. — P. 401—409.
301. *Sammon J. W.* An optimal discriminant plane // IEEE Trans. Comput. — 1970. — Vol. 19. — N 9.
302. *Schroeder A.* Analyse d'un mélange de distribution de probabilite de méme type // RSA. — 1976. — Vol. 24. — N1.
303. *Scott A. J., Symons M. J.* Clustering methods based on likelihood ratio criteria // Biometrics. — 1971. — Vol. 27. — P. 387.
304. *Semovskii S. V.* Linear discriminant analysis when the number of features is unbounded and the sample is finite. Тез. докл. I Всемири. конгр. общества матем. статистики и теории вероятн. им. Бернулли. — М.: Наука, 1986. — Т. 1. — Секции I— 19.— С. 210.
305. *Shepard R. N., Arabie P.* Additive clustering: presentation of similarities as combinations of discrete overlapping properties // Psycholog. Review. — 1979. — Vol. 86. — N 2. — P. 87—123.
306. *Shepard R. N., Carroll J. D.* Parametric representation of nonlinear data structures // Miltivariate analysis/Ed. Krishnaiah P. R. — N.-Y.: Acad. Press, 1966. — P. 561 — 592.
307. *Shepard R. N.* The analysis of proximites: multidimensional scaling with an unknown distance function // Psychometrika. — 1962. — Vol. 27. — N 2—3.

308. *Shepp L. A., Kruskal J. B.* Computerized tomography: the new medical X—ray technology // Amer. Math. Monthly. — 1978.— 75. — P. 420—439.
309. *Siegel J. B.* Software for microcomputers. — North — Holland, 1985.
310. *Silverman.* Some asymptotic properties of the probabilistic teacher // IEEE Inform. theory. — 1980. — Vol. 26. — N 2.
311. *Silvertein J. W.* The smallest eigenvalue of a large dimensional Wishart matrix // Ann. Probab. — 1985. — Vol. 13. — N 4. — P. 1364—1368.
312. *Skandinavisk aktuarietidskrift.* Haft 1. 1934.
313. *Smith A F.M., Spiegelhalter D. J.* Bayesian Approaches to Multiwariate Structure // J. R. Statist. Soc. — 1977. — B. — Vol 39. — N 1.
314. *Smith A. F. M., Makov M.* A quasi Bayes sequential procedure for mixtures // IRSS. — 1978. — B. — Vol. 40. — N1.
315. *Stuetzel W.* Plot windows // JASA. — 1987. — Vol. 82. — P. 466—475.
316. *Symons M. J.* Clustering criteria multivariate normal mixtures// Biometrics, — 1981. — Vol. 37.
317. *Takane Y., Young F. W., de Leeuw J.* Nonmetric individual differences multidimensional scaling: an alternative least squares method with optimal scaling features // Psychometrika. — 1977. — Vol. 42. — N1.
318. *Torgerson W. S.* Multidimensional Scaling. Theory and Method// Psychometrika. — 1952. — Vol. 17. — N 4.
319. *Takiyma R.* A two-level committee machine: a representation and a learning procedure for general piecewise linear discriminant function // Pattern Recognition. — 1981. — 13. — N3. — P. 269 — 274.
320. *Teicher H.* Identifiability of mixtures // Ann. Math. Statist. — 1961. — 32. — N 1. — P. 244—248.
321. *Teacher H.* Identifiability of finite mixtures // Ann. Math. Statist. — 1963. — 34. — N4 — P. 1265—1269.
322. *Truett J., Cornfield J., Kannel W. J.* A multivariate analysis of the risk of coronary heart disease in Framingham // Journ. Chron. Dis. — 1967 — 19. — P. 711—715.
323. *Tukey P A., Tukey J. W.* Graphical display of data sets in 3 or more dimensions· preparation: prechosen sequences of views // Inferpreting multivariate data / V. Barnet (ed). — Chichester: Wiley. — 1981. — P. 189—274.
324. *Usawa H.* Preference and rational choice in the theory function // Econometrika. — 1949. — 17. — N2. — P. 195.
325. *Van Campenhout J. N.* Topics in Measurment Selection // Handbook of Statistics // Krishnaiah P. R. (Ed). — Vol. 2: Classification, Pattern recognition and Reduction of dimensionality. — North-Holland. — 1982. — P. 793—803.
326. *Vasicek O.* A test of Normality Based on Sample Entropy // J.R. Statist. Soc. — 1976. — Vol. 38. — Ser. B. — P. 54—59.
327. *Yakowitz S. A., Spragins. J.* On the identifiability of finite mixtures // Ann. Math. Statist. — 1968. — 39. — N1. — P. 209—214.
328. *Yenyukov I. S.* Detecting outliers and clusters in multivariate data based on projection pursuit // Proceedings of the World Congress of Bernoulli Society, Tashkent, USSR, September

1986. — Utrecht: VHU Science Press. — 1987. — Vol. 2. — P. 137 — 141.
329. *Young F. W., Null C. H.* Multidimensional scaling of nominal data: the recovery of metric information with ALSCAL // Psychometrika. — 1978. — Vol. 43. — N 3.
330. *Willmott A. J., Grimshaw P.N.* Cluster analysis in social geography: Numerical taxonomy. — Ld., N. — Y.: Acad. Press., 1969.— P. 271—281.
331. *Wishart D.* An algorithm for hierarchical classification // Biometrics. — 1969 — 22 — P. 165 — 170.
332. *Wishart D.* Mode analysis. a generalization of nearest neighbour with reduce a chaining effect. Numerical Taxonomy. — Ld., N.-Y. Acad. Press., 1969. — P. 282—311.
333. *Wolfe J. H.* Pattern clustering by multivariate mixture analysis// Miltiv. Behav. Res. — 1969. — Vol. 5. — P. 329 —349.
334. *Wu.* On the convergence of the EM algorithm // Ann. Statist.— 1983. — Vol. 11. — N1.
335. *Zadeh L. A.* Fuzzy sets // Information and Control. — 1965. — Vol. 8. — P. 338—353.
336. *Zadeh L. A.* Outline of a new approach to the analysis of complex systems and decision process // IEEE Trans. on Systems, Man and Cybernetics. — 1973. — Vol. 3. — P. 28—44.

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

CONTENTS

Справочное издание

Сергей Арутюнович Айвазян,
Виктор Матвеевич Бухштабер,
Игорь Семенович Енюков,
Лев Дмитриевич Мешалкин

ПРИКЛАДНАЯ СТАТИСТИКА:
КЛАССИФИКАЦИЯ И СНИЖЕНИЕ
РАЗМЕРНОСТИ

Зав. редакцией *Р. А. Казьмина*
Редактор *Л. Н. Вылегжанина*
Мл. редакторы *В. Г. Крылова, Е. В. Гаврилова*
Худож. редактор *С. Л. Витте*
Техн. редакторы *Е. Д. Кузнецова, И. В. Завгородняя*
Корректоры *Г. В. Хлопцева, Г. А. Башарина, М. М. Виноградова, Н. П. Сперанская, Т. В. Рослякова*
Переплет художника *Н. А. Пашуро*

ИБ № 2182

Сдано в набор 15.06.88. Подписано в печать 6.01.89. А09236
Формат 84×108$^{1}/_{32}$. Бум. кн.-журн. Гарнитура литературная.
Печать офсетная. Усл. п. л. 31,92 Усл. кр.-отт. 31,92. Уч.-изд. л. 33,64. Тираж 15 000 экз. Заказ 291. Цена 2 руб.

Издательство «Финансы и статистика», 101000, Москва, ул. Чернышевского, 7.

Набрано в Московской типографии № 4 Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли, 129041, Москва, Б. Переяславская ул., д. 46.

Отпечатано в тип. им. Котлякова издательства «Финансы и статистика» Государственного комитета СССР по делам издательств, полиграфии и книжной торговли. 195273, Ленинград, ул. Руставели, 13.