Bĕnjamin S. Duran, Patrick L. Odĕll

CLUSTER ANALYSIS
A SURVEY

SPRINGER — VERLAG
BERLIN — HEIDELBERG — NEW YORK 1974

Б. Дюран и П. Оделл

КЛАСТЕРНЫЙ АНАЛИЗ

Перевод с английского *Е. З. Демиденко*
Научное редактирование и предисловие
А. Я. Боярского

МОСКВА „СТАТИСТИКА“ 1977

517.8
Д97

Дюран Б. и Оделл П.

Д97 Кластерный анализ. Пер. с англ. Е. З. Демиденко. Под ред. А. Я. Боярского. Предисловие А. Я. Боярского. М., «Статистика», 1977.

128 с. с ил.

Тема книги — обзор состояния теории и практики применения «кластерного анализа». Этот метод имеет все преимущества метода комбинационной группировки, но свободен от его главного недостатка — распыления материала, что открывает широкие перспективы применения рассматриваемого метода в статистическом анализе, в классификации объектов, в исследовании связей, типизации выборки и др. Книга отличается полнотой, доступностью и вместе с тем краткостью изложения.

Книга рассчитана на статистиков, экономистов, а также социологов, демографов, биологов и других специалистов.

517.8

Д $\frac{10803*\text{-}036}{008(01)\text{-}77}$ 40-77

* Второй индекс 10805.

О МЕТОДОЛОГИЧЕСКИХ ПРИНЦИПАХ И МНОГОМЕРНОМ АНАЛИЗЕ
(вместо предисловия)

Б. С. Ястремский, классифицируя задачи математической статистики, представил их в виде двух разрезов: статика — динамика и одномерные — многомерные задачи. Переход ко второй, более сложной ступени в обоих аспектах связан с положениями диалектики, требующей рассмотрения явлений в их развитии (динамике) и взаимосвязи, что и ведет к многомерным задачам. Не касаясь здесь первого аспекта, остановимся на втором. В область многомерных задач статистик вступает, как только он принимается за изучение совместной вариации двух признаков, т. е. их связи. Безразлично, идет ли при этом речь об аналитических группировках, комбинационной таблице по двум признакам с подсчетом числа случаев или о математических методах корреляции, дисперсионного анализа и т. п.

Развитие анализа в этом направлении приводит к рассмотрению взаимосвязи не пары признаков, а большего их числа. В области элементарных приемов это получает выражение в сложных комбинационных группировках с развитым сказуемым. В области математико-статистических методов речь идет о множественной корреляции, дисперсионном анализе зависимости от нескольких переменных и т. д.

На этом пути по мере углубления исследования рассматриваются связи все большего числа признаков. Но здесь возникают трудности технического характера. В свое время считались великолепным достижением исследования зависимостей, в которых количество одновременно учитываемых аргументов доводилось до десятка или больше. Старая литература изобилует примерами раздельного изучения зависимостей некоторого признака от каждого из большого числа других, а то и

просто перечнями влияющих признаков. Изучение их влияния в комплексе наталкивалось на два препятствия: технические трудности и ограниченность материала наблюдения. Второе чаще имеет место в естественнонаучных исследованиях и нередко оказывается снятым в социально-экономической статистике. В то время как экспериментатор должен получить значимые результаты из наблюдений над десятком кроликов, исследователь бюджета рабочей семьи располагает десятками тысяч наблюдений или социал-гигиенист — тысячами «историй болезни». Впрочем, при территориальной дифференциации или комбинации большого числа признаков не помогают и эти тысячи. В преодолении первого препятствия существенную роль играет новая вычислительная техника. Для ЭВМ получение, например, уравнения связи по сотне признаков не составляет проблемы.

Однако при старой технике роль качественного анализа (специфической логики специфического предмета) видна с первых же шагов выбора аргументов, поскольку драгоценные места для них в уравнении надо было расходовать с большой оглядкой, не допуская их траты на малосущественные связи. В условиях новой техники это соображение может показаться потерявшим значение. Для нее не так важно, если в числе сотни введенных в уравнение связи аргументов десяток окажется бесполезным, не влияющим на результативный признак: их введение не вытесняет из рассмотрения других, действительно важных.

Однако вскоре опыт показал, что голый эмпиризм остается тем, чем он был. Правда, если по его поводу Энгельс сказал, что и слепая свинья может найти свой желудь, то вооруженная современной электронной техникой она может найти целую горсть желудей. Но от этого в принципе ничего не меняется — эмпиризм остается эмпиризмом со своими возможностями и своей ограниченностью.

Сказанное мало интересует прагматически настроенных исследователей. Зато они ясно ощущают, что по мере возрастания размерности задачи все более теряется обозримость результатов. Закономерность распыляется на множество зачастую малозначащих связей.

С позиций эмпиризма сам переход от индивидуальных значений к обобщенным характеристикам есть реализация принципа «экономии мышления», а неимоверное возрастание числа этих характеристик при комплексном рассмотрении взаимосвязей многих признаков оказывается в полном противоречии с этим принципом. Исследователь снова стоит перед лицом огромной массы индивидуальных наблюдений. Так возникает задача обратного сведения множества характеристик к небольшому ряду обобщающих итогов, выражающему действительно существенное, закономерное для явления. Но пока каждый вовлеченный в анализ признак остается отдельным самостоятельным элементом, со своими характеристиками и линиями связи, число параметров, выражающих результаты обработки, не поддается уменьшению. Единственный путь к нему — либо в отсечении большинства признаков и возвращении к малоразмерным классическим задачам, либо в объединении признаков, в замене целых «гроздей» их одним, неизбежно искусственно построенным на их основе. Так появляется направление, получившее название «многомерный анализ». Его развитие и составляет новую ступень в истории математической статистики, которой отмечены последние десятилетия. Из предыдущего ясна и его связь с могущественной вычислительной техникой.

В многомерном анализе образовались разделы, которые, однако, не изолированы, а проникают и переходят один в другой. Это кластерный анализ (которому посвящена данная книга), таксономия, распознавание образов, метод главных компонент, факторный анализ. Наиболее ярко отражают черты многомерного анализа в классификации объектов кластерный анализ, а в исследовании связи — факторный анализ. Все эти разделы, закономерно обусловленные развитием математико-статистических методов и практики их применения, несомненно, несут в себе новые богатые возможности для решения многих познавательных задач, какими не располагают «классические» методы. В то же время их необходимо подвергнуть теоретическому анализу с позиций методологических принципов марксистско-ленинской теории познания. Основополагающим гносеологическим принципом статистической науки является примат качества объекта, его специфической логики. Усовершенствование формальных методов эмпирического исследования не может в какой-либо мере поколебать этот принцип. Методы многомерного анализа пока что представляются, однако, эмпиризмом, сбросившим всякие

оковы этой логики качества материального объекта (или, как теперь выражаются, «содержательного» анализа). Не случайно, например, факторный анализ возник в такой области, как психология, где для него почва особенно благоприятна, поскольку раскрытие внутренней логики в ней исключительно трудно. Задача в том, чтобы, используя приемы многомерного анализа, согласовать их теоретическую трактовку и применение с основным методологическим принципом статистического исследования. Без этого труд людей и дорогих машин легко окажется малоэффективным. Это часто видно, например, в факторном анализе, когда дело доходит до главной фазы исследования — выводов. Получив результаты вычислений, исследователю предстоит их «проинтерпретировать», иначе говоря, выяснить их смысл (экономический или иной). Выходит то, что должно быть в самом начале, выступает на сцену лишь в конце.

Сказанное можно вполне отнести и к кластерному анализу. Наиболее существенные его методологические черты сводятся к двум: образование единой меры, охватывающей ряд признаков, и чисто количественное решение вопроса о группировке объектов наблюдения.

Идея классификации по сочетанию ряда признаков не нуждается в аргументации. Как раз в одной из наиболее популярных задач кластерного анализа — группировке районов — она давно признана. Еще в 1920 г., анализируя «Связь между элементами крестьянского хозяйства в 1917 и 1919 годах» («Вестник статистики», 1920, с. 19—21), Б. С. Ястремский рассматривал 34 характеристики уездов, влиявшие на эту связь. Можно привести и другие примеры группировки территориальных единиц, по комплексу признаков, неизменно имевших место в задачах районирования. Но в кластерном анализе признаки объединяются с помощью некоторой «метрики» в один количественный показатель сходства (различия) группируемых объектов. Казалось бы, достаточно запустить в ЭВМ массив информации о них и подходящую программу классификации. На самом деле, однако, без предварительного анализа качества нельзя и приступить к делу. Уже в определении самого перечня признаков он неизбежно присутствует. Иной, быть может, скажет на это, что надо попросту ввести в ЭВМ весь материал наблюдения. Но ведь кто-то на основании чего-то составил и саму программу наблюдения.

Стоит себе отдать в этом ясный отчет и все становится на свое место, настолько, что в этой части кластерный анализ оказывается сродни идеям таких классических исследований, как ленинские группировки крестьянских хозяйств, выявившие два класса капиталистической экономики на полюсах и еще не размытую дифференциацией середину. Капиталистическая верхушка — это хозяйства эксплуататоров. Но эксплуатация могла осуществляться в разных формах: наем работников, «прокат» инвентаря, займы и т. д. Для этого хозяйство должно было чем-то располагать: землей, инвентарем, деньгами, мастерской или торговлей. Как видим, принадлежность к этой группе внешне могла получать отражение в ряде признаков. Отсюда ленинское требование изучать совокупность признаков, давать по ним группировку не параллельную, а комбинационную. При этом важность признаков зависит от особенностей района: в земледельческом—это прежде всего площадь посева, в животноводческом — численность скота и т. д. Легко видеть, что как сама задача, так и необходимость учета в ее решении ряда признаков и выбора этих признаков — все это продиктовано качественным анализом и, как всегда в статистике, через него тесными узами связано с целью исследования. Наверно, если бы цель состояла не в анализе классовой дифференциации, а в анализе уровня культуры, ведущее место заняли бы грамотность, число лет обучения, наличие в доме книг и т. п., а посевная площадь или число лошадей заняли бы место вспомогательной информации для выявления связи между уровнем культуры и социально-экономическим фактором.

Коль скоро признаки отобраны, может быть оправданным и подход кластерного анализа, но не как чисто эмпирического, а основанного на правильных методологических принципах. Так, именно из качественного анализа вытекает, что в составе капиталистической верхушки могли быть хозяйства, обрабатывающие большой земельный массив и без большого числа лошадей, и одновременно хозяйства без больших посевов, но богатые живым и мертвым инвентарем или имеющие торговлю и т. д. Следуя указаниям качественного анализа, их надо объединить в одну группу.

В кластерном анализе группировочные признаки подвергаются объединению с помощью некоторой «метри-

ки» — евклидова расстояния или иной. Но здесь возникает самое настоящее embarras de richesse, затруднение от изобилия. Метрик оказывается много и число их возрастает. Какой отдать предпочтение? Кроме того, в частности, в евклидовой результат зависит от масштаба, от выбранных единиц измерения, например, будет ли один признак измеряться в метрах, другой в килограммах или первый в сантиметрах, а второй в тоннах. Это обстоятельство вскользь отмечает и автор данной книги. Правда, есть способ выйти из затруднения путем нормирования признаков. Но нельзя доказать, что для всех признаков одно квадратическое отклонение одинаково значимо.

Вопрос о выборе метрики и масштабов имеет различное содержание в зависимости от целей. Если группировки различаются на «типологические» и «аналитические» (не настаиваем на этой терминологии), то же самое не может не относиться и к кластеризации. Между тем в литературе это игнорируется. Более того, кластеры выдаются обычно за «типы», что должно в какой-то мере подчеркивать их существенное различие, чуть ли не качественное.

Если речь идет о качестве в подлинном смысле слова, то ни метрики, ни масштабы не произвольны. Так, для выделения верхней группы крестьянских хозяйств надо было учитывать ряд признаков, но так, чтобы их сочетание давало основание для причисления хозяйства к этой группе. Критерием могла бы быть совокупная величина возможного дохода. Поставив для нее некоторую нижнюю границу, отвечающую возможности основывать хозяйство на прибавочной стоимости без участия в нем личным трудом, мы бы получили объективный критерий для масштабов, да и для метрики. В одних случаях такой подход не слишком труден. Например, мощность тракторов и число лошадей сравнительно легко бы поддавались соизмерению. В других задача гораздо труднее. Но сказанное должно служить ориентиром во всех случаях типологической кластеризации (если можно так выразиться).

Другое дело формально-количественная кластеризация. Ее цели скромнее: представить в сжатом виде массив информации с его многомерностью, но так, чтобы потеря информации не была чрезмерной. Здесь нет жестких объективных требований и решение может быть

различным. То же можно сказать по поводу любой «аналитической» группировки. Общность вопроса вытекает уже из того, что группировка по одному признаку и кластеризация по ряду признаков приводятся друг к другу. Число соединяемых при кластеризации признаков может быть равным и единице. Это приводит задачу группировки по одному признаку к кластеризации. С другой стороны, используемая при объединении признаков метрика сводит их к одному признаку и далее разбиение на кластеры равнозначно группировке по этому признаку. О путях формализации последней задачи уже немало сказано в литературе. Внедрение ЭВМ и перевод обработки информации на индустриальные рельсы не может оставить на субъективный произвол число и границы интервалов группировки. Значит, неизбежно применение в этом некоторого формального стандарта. Однако таких стандартов может быть несколько: разбиение по децилям, по квадратическим отклонениям, по максимуму «локального расстояния», по относительному расстоянию, по внутрикластерному коэффициенту вариации и т. д. Индустриальный подход, таким образом, не исключает инициативы исследователя, его выбора. Но этот выбор теперь будет состоять в выборе между несколькими стандартами, для которых имеются машинные программы. Это несколько ограничивает исследователя, но дает возможность гораздо большей сравнимости разных группировок и их быстрого получения.

Выходит, что методы кластеризации нужны при внедрении ЭВМ даже для решения задачи простой группировки. Поскольку в ней нет качественного критерия, все сводится к образованию групп по количественному сходству. А в такой постановке машина с помощью той или иной стандартной программы может с ней справиться лучше.

Из всего сказанного ясно, что по отношению к кластерному анализу, как и к другим частям многомерного анализа, необходимо, во-первых, хорошо изучить теорию и имеющуюся практику применения, во-вторых, на основе этого и все увеличивающегося нового опыта применения глубоко осмыслить его технику с позиций общих методологических принципов статистической науки.

В достижении первой цели предлагаемая книга представляет большую ценность, так как в ней читатель най-

дет богатый и в то же время сжато изложенный материал, образующий в целом прекрасный обзор теории кластерного анализа и ряда его приложений. Именно эту цель и ставили перед собой авторы и они прекрасно справились с делом. Что касается второго, то эту задачу мы здесь, конечно, могли только поставить. Она должна быть решена не столько математиками, сколько статистиками, экономистами и представителями других конкретных областей применения, не без участия философов.

В целом же книга заслуживает высокой оценки не только как монография, но и как пособие учебного характера. Хотя некоторые ее места воспринимаются не без известного труда, в целом она отличается от многих других книг по этой или примыкающим проблемам ясностью и доступностью изложения. Ее появление на русском языке несомненно принесет большую пользу советским специалистам и всем интересующимся статистической наукой.

Редактор взял на себя смелость исправить некоторые явные опечатки оригинала.

А. Я. БОЯРСКИЙ

ПРЕДИСЛОВИЕ

За последнее тридцатилетие в области кластерного анализа была проделана большая работа, причем значительная ее часть была проведена после 1960 г. Основное содержание этой книги было опубликовано в различных журналах, в том числе прикладного характера, однако до сих пор этот материал не был собран воедино.

Цель данной монографии заключается в том, чтобы объединить разрозненные статьи в виде краткого обзора по кластерному анализу.

Мы надеемся, что эта книга позволит читателю быстро ознакомиться с проблемами кластерного анализа и другими смежными вопросами.

По этой причине многие детали были опущены. Это же касается иллюстрирующих примеров. Большинство работ, на которые мы ссылаемся, содержат примеры применения кластерного анализа, поэтому читатель может воспользоваться ими для получения дополнительной информации по специальным вопросам. Мы постарались включить в библиографию все работы, которые сыграли какую-либо роль в развитии «теории» кластерного анализа. Этот список содержит также работы прикладного характера, однако наша библиография все же далеко не полная.

Эта монография была написана в значительной мере под влиянием работ многих исследователей в данной области; это в первую очередь относится к работам Хартигана, Уишарта, Брайена, Дженсена, Вайнода и Рао.

Изложение некоторых частей книги основано на исследовании, выполненном при поддержке Центра пилотируемых космических кораблей НАСА (отдел наземного наблюдения, контракт NAS 9—12775).

ГЛАВА 1

ПРОБЛЕМА КЛАСТЕРНОГО АНАЛИЗА. ОСНОВНЫЕ ИДЕИ

1.1. Основные обозначения и определения

Методы кластерного анализа можно применять в различных ситуациях, встречающихся в исследованиях как научных, так и чисто прикладного характера. В этой главе мы не будем останавливаться на специфических особенностях приложения кластерного анализа в тех или иных областях, а рассмотрим его технику с общих, может быть, несколько абстрактных позиций. Интересные, с нашей точки зрения, приложения дадим в другой главе.

Пусть множество $I=\{I_1, I_2, \ldots, I_n\}$ обозначает n объектов (индивидов), принадлежащих некоторой популяции π_I. Предположим также, что существует некоторое множество *наблюдаемых* показателей или характеристик $C=(C_1, C_2, \ldots, C_p)^T$*, которыми обладает каждый индивид из I. Наблюдаемые характеристики могут быть как *количественными,* так и *качественными;* однако основная часть нашего рассмотрения будет посвящена количественным данным, которые иногда будем называть *измерениями.* Результат измерения i-й характеристики I_j объекта будем обозначать символом x_{ij}, а вектор $X_j=[x_{ij}]$ размерности $p\times 1$ будет отвечать каждому ряду измерений (для j-го индивида). Таким образом, для множества индивидов I исследователь располагает множеством векторов измерений $X=\{X_1, X_2, \ldots, X_n\}$, которые описывают множество I. Отметим, что множество X может быть представлено как n точек в p-мерном евклидовом пространстве E_p.

* Здесь ^{T} — знак транспонирования, который в русской литературе обычно обозначается штрихом. — *Примеч. ред.*

1.2. Задача кластерного анализа

Пусть m — целое число, меньшее, чем n. Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся в множестве X, разбить множество объектов I на m кластеров (подмножеств) $\pi_1, \pi_2, \ldots, \pi_m$ так, чтобы каждый объект I_i принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были *сходными,* в то время как объекты, принадлежащие разным кластерам, были разнородными (*несходными*)*.

Решением задачи кластерного анализа является разбиение, удовлетворяющее некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок. Этот функционал часто называют *целевой функцией.* Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонений (см. параграф 1.5). В качестве примера рассмотрим $n=8$ объектов, обладающих одной характеристикой (т. е. $p=1$); результаты измерения пусть представляют собой множество $X=\{3, 4, 7, 4, 3, 3, 4, 4\}$. Сумма квадратов отклонений вычисляется по формуле

$$W=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2,$$

где x_i представляет собой измерение i-го объекта. Для нашего примера, содержащего 8 объектов, получим:

$$\sum_{i=1}^{8}x_i^2-\frac{1}{8}\left(\sum_{i=1}^{8}x_i\right)^2=140-128=12.$$

Если множество X разбить на три группы: $G_1=\{3, 3, 3\}$, $G_2=\{4, 4, 4, 4\}$ и $G_3=\{7\}$, то все внутригрупповые суммы квадратов отклонений будут равны нулю:

$$W_1+W_2+W_3=0+0+0=0,$$

* Приведем следующий пример задачи кластерного анализа. В качестве I рассмотрим n стран, каждую из которых характеризуем валовым национальным продуктом на душу населения (C_1), личным потреблением на душу населения (C_2), душевым потреблением электроэнергии (C_3) и т. п. Тогда X_1 (вектор измерений) представляет собой набор указанных характеристик для первой страны; X_2 — для второй и т. д. Задача заключается в том, чтобы разбить страны по уровню развития. — *Примеч. пер.*

где W_i обозначает сумму квадратов, соответствующую группе G_i. Оптимальное значение для этого примера равно нулю при условии, что ведется разбиение на три группы. В общем случае следует рассматривать значение целевой функции в сочетании с желаемым числом групп. Далее будут определены различные виды целевых функций, многие из которых могут быть записаны в универсальной и общей форме.

Очевидно, для того чтобы «решить» задачу кластерного анализа, необходимо количественно определить понятия сходства и разнородности. Что означает «два объекта I_j и I_k различны»? Задача была бы решена, если бы i-й и j-й объекты попадали в один и тот же кластер всякий раз, когда расстояние (отдаленность) между соответствующими точками X_i и X_j было бы «достаточно малым», и, наоборот, попадали в разные кластеры, если бы расстояние между точками X_i и X_j было бы, «достаточно большим». Таким образом, для нашей цели следует рассмотреть понятие расстояния между точками X_i и X_j из E_p с абстрактных позиций.

1.3. Функции расстояния

Определение 1.1. Неотрицательная вещественнозначная функция $d(X_i, X_j)$ называется функцией расстояния (метрикой), если:

а) $d(X_i, X_j) \geqslant 0$ для всех X_i и X_j из E_p;

б) $d(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$;

в) $d(X_i, X_j) = d(X_j, X_i)$;

г) $d(X_i, X_j) \leqslant d(X_i, X_k) + d(X_k, X_j)$, где X_i, X_j и X_k — любые три вектора из E_p.

Значение $d(X_i, X_j)$ для заданных X_i и X_j называется расстоянием между X_i и X_j и эквивалентно расстоянию между I_i и I_j соответственно выбранным характеристикам $(C_1, C_2, \ldots, C_p)^T$.

В таблице 1.1 приводятся примеры некоторых наиболее употребительных функций расстояния.

Евклидова метрика очень популярна и наиболее употребительна. Метрика l_1 абсолютных значений наиболее простая с вычислительной точки зрения. Сюпремум-норма также легко вычисляется и включает в себя процедуру упорядочивания. l_p-норма охватывает функции расстояния 1, 2 и 3, соответственно $p = 2{,}1$ и ∞.

Расстояние Махаланобиса часто называют обобщенным евклидовым расстоянием. W^{-1} обычно обозначает

Таблица 1.1. **Некоторые функции расстояния**

Название	Формула
1. Евклидово расстояние	$d_2(X_i, X_j) = \left[\sum_{k=1}^{p} (x_{ki} - x_{kj})^2\right]^{1/2}$
2. l_1-норма	$d_1(X_i, X_j) = \left[\sum_{k=1}^{p} \lvert x_{ki} - x_{kj}\rvert\right]$
3. Сюпремум-норма	$d_\infty(X_i, X_j) = \sup_{k=1,2,\ldots,p} \{\lvert x_{ki} - x_{kj}\rvert\}$
4. l_p-норма	$d_p(X_i, X_j) = \left[\sum_{k=1}^{p} \lvert x_{ki} - x_{kj}\rvert^p\right]^{1/p}$
5. Махаланобиса [252]	$D^2(X_i, X_j) = (X_i - X_j)^T W^{-1} (X_i - X_j)$

матрицу, обратную к матрице рассеяния (см. параграф 1.5). Расстояние Махаланобиса инвариантно относительно невырожденных линейных преобразований. Рассмотрим преобразование $Y = BX$. Тогда

$$D^2(Y_i, Y_j) = (Y_i - Y_j)^T W_y^{-1} (Y_i - Y_j) =$$

$$= (BX_i - BX_j)^T W_y^{-1} (BX_i - BX_j) =$$

$$= (X_i - X_j)^T B^T W_y^{-1} B (X_i - X_j) =$$

$$= (X_i - X_j)^T B^T (B W_x B^T)^{-1} B (X_i - X_j) =$$

$$= (X_i - X_j)^T W_x^{-1} (X_i - X_j) = D^2(X_i, X_j).$$

Существуют другие, эвристические, меры отдаленности, не являющиеся расстояниями с точки зрения определения 1.1, которые, однако, также применяются на практике. Например, мера Джеффриса — Матуситы [81], [82], [259], которая определяется по формуле

$$M = \left[\sum_{k=1}^{p} (\sqrt{x_{ki}} - \sqrt{x_{kj}})^2\right]^{1/2}, \qquad (1.1)$$

и другая мера, известная под названием «коэффициент дивергенции» [55]

$$CD = \left\{\frac{1}{p}\sum_{k=1}^{p}\left(\frac{x_{ki} - x_{kj}}{x_{ki} + x_{kj}}\right)^2\right\}^{1/2}.$$

Мера Джеффриса — Матуситы первоначально была введена в качестве расстояния между двумя функциями плотностей вероятности, однако в форме (1.1) она может быть применена и как мера расстояния между парой векторов. В первоначальном применении коэффициента дивергенции x были действительными средними $\bar{x}$ и рассматривались как расстояние между выборочными средними двух выборок.

Следующая теорема позволяет упорядочить функции расстояния, определяемые по l_p-норме.

Теорема 1.1. Неравенство $d_h(X_i, X_j) \leqslant d_m(X_i, X_j)$ выполняется для всех X_i и X_j из E_p тогда и только тогда, когда $h \geqslant m$*.

Напомним, что из определения расстояния следует, что для $X_i = X_j$, $d_p(X_i, X_j) = 0$.

1.4. Меры сходства

n измерений $X_1, X_2, \ldots, X_n$ могут быть представлены в виде матрицы данных размером $p \times n$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & & \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{pmatrix} = (X_1, X_2, \ldots, X_n).$$

Аналогичным образом расстояния между парами векторов $d(X_i, X_j)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & d_{12} & \ldots & d_{1n} \\ d_{21} & 0 & \ldots & d_{2n} \\ \vdots & \vdots & & \\ d_{n1} & d_{n2} & \ldots & 0 \end{pmatrix}.$$

Заметим, что диагональные элементы $d_{ii} = 0$ для $i = 1, 2, \ldots, n$.

Понятием, противоположным расстоянию между X_i и X_j, является понятие сходства между двумя объектами I_i и I_j.

---

* Ср. мажорантность средних в применении к степенным средним. — *Примеч. ред.*

Определение 1.2. Неотрицательная вещественная функция $s(X_i, X_j) = s_{ij}$ называется мерой сходства, если:

1) $0 \leqslant s(X_i, X_j) < 1$ для $X_i \neq X_j$;
2) $s(X_i, X_i) = 1$;
3) $s(X_i, X_j) = s(X_j, X_i)$.

Пары значений мер сходства можно объединить в матрицу сходства:

$$S = \begin{pmatrix} 1 & s_{12} & \ldots & s_{1n} \\ s_{21} & 1 & \ldots & s_{2n} \\ \vdots & \vdots & & \\ s_{n1} & s_{n2} & \ldots & 1 \end{pmatrix}.$$

Величину s_{ij} будем просто называть коэффициентом сходства. Если каждый вектор измерения X_i состоит из нулей и единиц, эту величину называют коэффициентом ассоциации, или парным коэффициентом сопряженности.

Существует несколько видов коэффициентов ассоциации, значения которых лежат в пределах от -1 до $+1$. К этой группе принадлежит фи-коэффициент, известный также под названием «четырехпольный коэффициент корреляции». В дальнейшем мы остановимся только на коэффициентах, удовлетворяющих определению 1.2.

Предположим, что каждый вектор наблюдений содержит только нули и единицы, т. е. бинарные данные. Для заданных векторов X_i и X_j обозначим через n_{IJ} число характеристик, которые соответствуют единицам в векторах X_i и X_J; через n_{ij} — число характеристик, соответствующих нулям в этих векторах; через n_{iJ} — число характеристик, дающих нуль в X_i и единицу в X_J; сходным образом определяется n_{Ij}. Таким образом, $n_J = n_{IJ} + n_{iJ}$ есть число единиц в X_j, а $n_j = n_{Ij} + n_{ij}$ — число нулей в X_j. В табл. 1.2 приводятся примеры коэффициентов сходства, выраженных в терминах определенных выше величин. Обсуждение коэффициентов сходства табл. 1.2, а также другие коэффициенты читатель найдет в работе [336].

Статистики постоянно пользуются мерой линейного сходства, называемой коэффициентом корреляции, который обычно обозначается r_{ij} и вычисляется по формуле

$$r_{ij} = \left[\sum_{k=1}^{p} x_{ki} x_{kj}\right] \Big/ \left[\sum_{k=1}^{p} x_{ki}^2 \sum_{k=1}^{p} x_{kj}^2\right]^{1/2}. \qquad (1.2)$$

Таблица 1.2. **Коэффициенты сходства для бинарных данных**

Коэффициент	Ссылка
$\frac{n_{IJ}}{n_{IJ}+n_{Ij}+n_{iJ}}$	[173], [328]
$\frac{n_{IJ}+n_{ij}}{p}$	[334]
$\frac{n_{IJ}}{p}$	[303]
$\frac{2n_{IJ}}{2n_{IJ}+n_{Ij}+n_{iJ}}$	[82], [338]
$\frac{2(n_{IJ}+n_{ij})}{p+n_{IJ}+n_{ij}}$	
$\frac{n_{IJ}}{n_{IJ}+2(n_{Ij}+n_{iJ})}$	
$\frac{n_{IJ}+n_{ij}}{p+n_{Ij}+n_{iJ}}$	[294]

В формуле (1.2) предполагается, что $\sum_{k=1}^{p} x_{ki} = \sum_{k=1}^{p} x_{kj} = 0$.

Коэффициент r_{ij} занимает важное место в статистике и употребляется, зачастую ошибочно, почти каждым. Важно подчеркнуть, что если X_i и X_j рассматривать как координаты двух точек в пространстве E_p, являющиеся концами двух векторов с началом в начале координат, то [7]

$$r_{ij} = \cos\theta, \qquad (1.3)$$

где θ — угол между этими двумя векторами. Поэтому, как следует из уравнения (1.3), $-1 \leqslant r_{ij} \leqslant 1$. Будем говорить, что объекты I_i и I_j сходны положительным образом (положительно), если r_{ij} «близок» к 1, отрицательно сходны, если r_{ij} «близок» к -1, не сходны, если r_{ij} «близок» к нулю. Заметим, что r_{ij} не является функцией сходства с точки зрения определения 1.2*.

* Не выполняется аксиома 1. — *Примеч. пер.*

Лемма 1.1. Коэффициент корреляции $r_{ij}=1$ тогда и только тогда, когда $X_i = kX_j$, где k — неотрицательное число.

Доказательство этой леммы следует непосредственно из формулы (1.2). Заметим, что две точки X_1 и X_2 могут быть сравнительно далекими друг от друга и в то же время сходство, измеряемое r_{ij}, может оказаться равным 1. Рассмотрим, в частности, следующий пример (график с числовыми значениями представлен на рис. 1).

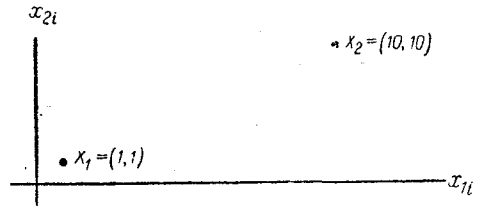


Рис. 1. Две точки в E_2

Пользуясь метриками (1), (2) и (3) табл. 1.1 и r_{ij} (уравнение (1.2)), найдем:

$$d_2(X_1, X_2) = [(10-1)^2 + (10-1)^2]^{1/2} = 9\sqrt{2};$$
$$d_1(X_1, X_2) = [|10-1| + |10-1|] = 18;$$
$$d_\infty(X_1, X_2) = \sup[|10-1|, |10-1|] = 9;$$
$$r_{12} = 1.$$

Заметим, что хотя $X_1 \neq X_2$, $r_{ij}=1$, т. е. объекты I_1 и I_2 с точки зрения такого критерия будут считаться сходными. Заметим также, что

$$d_\infty(X_1, X_2) < d_2(X_1, X_2) < d_1(X_1, X_2),$$

что лишний раз иллюстрирует теорему 1.1 из параграфа 1.3.

Важно заметить, что, выбирая соответствующее преобразование, можно исходя из различных мер расстояния, приведенных в параграфе 1.3, построить соответствующие меры сходства*. Поэтому если предпочтитель-

* Например, можно положить $s_2(X_i, X_j) = 1/(1+d_2(X_i, X_j))$, где d_2 — евклидово расстояние. Легко проверить, что все свойства меры сходства для s_2 выполняются (определение 1.2). — *Примеч. пер.*

нее работать с мерами сходства, то необходим соответствующий переход.

Воспользуемся теперь введенным понятием расстояния для вычисления меры рассеяния или разнородности множества объектов $I=\{I_1, \ldots, I_n\}$.

Определение 1.3. Пусть $X=\{X_1, X_2, \ldots, X_n\}$ обозначает множество наблюдений, произведенных над множеством объектов $I=\{I_1, I_2, \ldots, I_n\}$. Величина

$$s_d = 1/2 \sum_{i=1}^{n} \sum_{j=1}^{n} d(X_i, X_j) \qquad (1.4)$$

называется *общим* рассеянием, соответствующим данной функции расстояния $d(X_i, X_j)$.

Определение 1.4. Величина $\bar{s}_d = s_d/N_d$, где $N_d = (n^2-n)/2$ называется средним рассеянием множества I.

Обоснование определений 1.3 и 1.4 следует из рассмотрения матрицы расстояний $D=\{d_{ij}=d(X_i, X_j)\}$ с учетом того, что, во-первых, для всех i

$$d_{ii}=d(X_i, X_i)=0,$$

а, во-вторых, из $d(X_i, X_j)=d(X_j, X_i)$ следует $d_{ij}=d_{ji}$ для всех $i \neq j=1, 2, \ldots, n$.

Отсюда величина s_d представляет собой сумму n^2 расстояний, из которых n равны нулю, и $(n^2-n)/2$, вообще говоря, различны и неотрицательны. Поэтому $\bar{s}_d$ есть арифметическое среднее ненулевых расстояний между парами элементов из X, или, что то же, из I. Матрица D является компактной записью расстояний всех пар элементов из множества I.

Статистики применяют аналогичную меру рассеяния (см., например, Уилкс [391 с. 591—614]).

Определение 1.5. Матрица $p \times p$

$$S_x = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \qquad (1.5)$$

называется матрицей рассеяния множества X, причем

$$\bar{X} = \sum_{i=1}^{n} X_i/n \qquad (1.6)$$

есть вектор $p \times 1$ арифметических средних.

Матрицу S_x также иногда называют *матрицей суммы квадратов.*

Определение 1.6. След матрицы S_x называется *статистическим рассеянием* множества X и обозначается

$$s_t = \operatorname{tr} S_x = \sum_{i=1}^{n} \sum_{k=1}^{p} (X_{ki} - \bar{X}_k)^2 = \sum_{i=1}^{n} (X_i - \bar{X})^T (X_i - \bar{X}).$$

Мера s_t равна сумме квадратов расстояний n точек от средних по группе $\bar{X}$ и представляет собой сумму (внутреннюю по группе) квадратов отклонений. Можно показать, что

$$s_t = \frac{1}{n} \sum_{\substack{i=1 \\ i<j}}^{n} \sum_{j=1}^{n} (X_i - X_j)^T (X_i - X_j) = \frac{1}{n} \sum_{\substack{i=1 \\ i<j}}^{n} \sum_{j=1}^{n} d^2(X_i, X_j). \qquad (1.7)$$

Таким образом, когда оперируют следом $\operatorname{tr} S_x$, имеют в виду расстояние в евклидовом смысле.

Определение 1.7. Определитель $|S_x|$ матрицы S_x называется статистическим рассеянием, соответствующим определителю, и обозначается $s_D = |S_x|$*.

Матрица коэффициентов корреляций $R=\{r_{ij}\}$ может быть получена из матрицы $S_x=\{s_{ij}\}$, определенной уравнением (1.5). Найдем диагональную матрицу $[\operatorname{Dia} S_x]=\{s_{11}, s_{22}, \ldots, s_{pp}\}$ и $[\operatorname{Dia} S_x]^{1/2}=\{s_{11}^{1/2}, s_{22}^{1/2}, \ldots, s_{pp}^{1/2}\}$. Тогда

$$R = [\operatorname{Dia} S_x]^{1/2} S_x [\operatorname{Dia} S_x]^{1/2}. \qquad (1.8)$$

Лемма 1.2. $S_x=0$ (нулевой матрице) тогда и только тогда, когда $X_1=X_2=\ldots=X_k$ и $X_{k+1}=\ldots=X_n=0$ для некоторого $k \leqslant n$.

1.5. Расстояние между кластерами и их сходство

Как мы увидим позднее, многие процедуры при кластеризации совершаются ступенчато. Это означает, что два наиболее близко расположенных объекта I_1 и I_2 объединяются и рассматриваются как один кластер. Это приводит к тому, что число объектов уменьшается и становится равным $n-1$, причем один кластер будет содержать два объекта, а $n-2$ остальных по одному. Процесс можно повторять до тех пор, пока все объекты не сгруппируются в один кластер. В рассмотренной по-

* В статистике $|S_x|$ также называют обобщенной дисперсией. — *Примеч. пер.*

следовательной процедуре пользуются интуитивным представлением о расстоянии между объектом и кластером и расстоянии между двумя кластерами.

Неотъемлемой частью задачи кластерного анализа является понятие *оптимального критерия (целевой функции)*, которое позволяет установить, когда достигается желательное разбиение. Для введения подобного критерия необходимо найти меру внутренней *однородности* кластера и меру *разнородности* кластеров между собой.

Пусть $I=\{I_1, I_2, \ldots, I_n\}$ и $J=\{J_1, J_2, \ldots, J_{n_2}\}$ обозначают два кластера объектов, принадлежащих некоторой популяции π. Пусть $C=(C_1, C_2, \ldots, C_p)^T$ будет множеством характеристик, которые генерируют два множества измерений $X=\{X_1, X_2, \ldots, X_{n_1}\}$ и $Y=\{Y_1, Y_2, \ldots, Y_{n_2}\}$, соответствующие I и J.

Определение 1.8. Обозначим через $D=\{d(X_i, Y_j), i=2, \ldots, n_1; j=1, 2, \ldots, n_2\}$ множество всех расстояний. Величину

$$D_1(I, J) = \min d(X_i, Y_j),$$
$$i=1, \ldots, n_1,$$
$$j=1, \ldots, n_2$$

будем называть *минимальным локальным расстоянием* (nearest neighbor distance) [395] между кластерами I и J, соответствующим данной функции расстояния d.

Определение 1.9. Пусть $D=\{d(X_i, Y_j)\}$ определено так же, как и в определении 1.8. Тогда

$$D_2=\max d(X_i, Y_j),$$
$$i=1, \ldots, n_1,$$
$$j=1, \ldots, n_2$$

назовем *максимальным локальным расстоянием* (furthest neighbor distance) [234] между I и J.

Определение 1.10. Величина

$$D_3=\sum_{j=1}^{n_2}\sum_{i=1}^{n_1} d(X_i, Y_j)/n_1 n_2$$

есть *среднее расстояние* [225] между I и J, соответствующее данной функции расстояния d.

При оперировании понятием статистического рассеяния иногда пользуются следующей мерой расстояния между кластерами I и J.

Определение 1.11. Величину

$$D_4=\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})^T(\overline{X}-\overline{Y}),$$

где

$$\overline{X}=\sum_{i=1}^{n_1} x_i/n_1, \quad \overline{Y}=\sum_{i=1}^{n_2} y_i/n_2,$$

называют *статистическим расстоянием между кластерами* I и J*.

Меру D_4, очевидно, можно обосновать следующим образом. Рассмотрим два кластера I и J, которые в свою очередь составляют кластер K, где $K=I \cup J$ (значок $\cup$ означает объединение); тогда по формуле (1.5)

$$S_K=\sum_{i=1}^{n_1}(X_i-M)(X_i-M)^T+\sum_{i=1}^{n_2}(Y_i-M)(Y_i-M)^T,$$

где $M=(\sum_{i=1}^{n_1} X_i+\sum_{i=1}^{n_2} Y_i)/(n_1+n_2)$.

Поэтому

$$S_K=\sum_{i=1}^{n_1}(X_i-\overline{X}+\overline{X}-M)(X_i-\overline{X}+\overline{X}-M)^T+$$
$$+\sum_{i=1}^{n_2}(Y_i-\overline{Y}+\overline{Y}-M)(Y_i-\overline{Y}+\overline{Y}-M)^T=$$
$$=\sum_{i=1}^{n_1}(X_i-\overline{X})(X_i-\overline{X})^T+\sum_{i=1}^{n_1}(\overline{X}-M)(\overline{X}-M)^T+$$
$$+\sum_{i=1}^{n_2}(Y_i-\overline{Y})(Y_i-\overline{Y})^T+\sum_{i=1}^{n_2}(\overline{Y}-M)(\overline{Y}-M)^T,$$

поскольку $\sum_{i=1}^{n_1}(X_i-\overline{X})(\overline{X}-M)^T=\sum_{i=1}^{n_2}(Y_i-\overline{Y})(\overline{Y}-M)^T=0$.

Заметим, что

$$M=(n_1\overline{X}+n_2\overline{Y})/(n_1+n_2),$$

поэтому

$$\sum_{i=1}^{n_1}(\overline{X}-M)(\overline{X}-M)^T=\frac{n_1 n_2^2}{(n_1+n_2)^2}(\overline{X}-\overline{Y})(\overline{X}-\overline{Y})^T$$

* Легко видеть, что D_4 пропорциональна квадрату расстояния между «центрами» рассеяния множеств X и Y. — *Примеч. пер.*

и

$$\sum_{i=1}^{n_2}(\overline{Y}-M)(\overline{Y}-M)^T=\frac{n_1^2 n_2}{(n_1+n_2)^2}(\overline{X}-\overline{Y})(\overline{X}-\overline{Y})^T .$$

Окончательно

$$\sum_{i=1}^{n_1}(\overline{X}-M)(\overline{X}-M)^T+\sum_{i=1}^{n_2}(\overline{Y}-M)(\overline{Y}-M)^T=$$

$$=\left[\frac{n_1 n_2^2}{(n_1+n_2)^2}+\frac{n_1^2 n_2}{(n_1+n_2)^2}\right](\overline{X}-\overline{Y})(\overline{X}-\overline{Y})^T=$$

$$=\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})(\overline{X}-\overline{Y})^T .$$

Последнее выражение будем называть *матрицей межгруппового рассеяния.*

В результате получим:

$$S_K=S_I+S_J+\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})(\overline{X}-\overline{Y})^T, \qquad (1.9)$$

где S_I и S_J обозначают внутригрупповое рассеяние I и J.

Матрицу

$$\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})(\overline{X}-\overline{Y})^T \qquad (1.10)$$

назовем матрицей межгруппового рассеяния, а след этой матрицы

$$\mathrm{tr}\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})(\overline{X}-\overline{Y})^T=\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})^T(\overline{X}-\overline{Y})$$

— статистическим расстоянием между кластерами I и J. След матрицы (1.10) называют внутригрупповой суммой квадратов (ВСК). При объединении I и J в один кластер K, очевидно, ВСК возрастает.

Уравнение (1.9) статистики интерпретируют следующим образом: «*общая* сумма квадратов равна *внутригрупповой* сумме квадратов плюс *межгрупповая* сумма квадратов». Сумма S_I+S_J есть «внутригрупповая сумма квадратов», а выражение (1.10) представляет «межгрупповую сумму квадратов», записанную в матричной форме. Подобным образом можно было бы построить новые меры расстояния между кластерами, воспользовавшись другими функциями расстояния, рассмотренными в параграфе 1.3*.

* Напомним, что введенные здесь расстояния между кластерами опираются на евклидову метрику. — *Примеч. пер.*

Рассмотрим теперь несколько иной подход к проблеме измерения расстояния между кластерами. Предположим, что каждый кластер представляет собой выборку из некоторой генеральной совокупности (популяции). Обозначим через f и g функции плотности вероятности, соответствующие кластерам I и J. Уоккер и Лангриб [383] рассматривают различные многомерные формы мер расстояния и их метрические свойства. Их результаты сведены в табл. 1.3, где C обозначает класс всех p-мерных абсолютно непрерывных функций распределения, MVN — класс многомерных нормальных распределений, MVN_Σ — класс многомерных нормальных распределений с одинаковыми матрицами ковариаций. В таблице также приводятся метрические свойства мер расстояния соответственно для трех классов функций распределения.

Эти меры межкластерного расстояния могут оказаться весьма полезными в случае нормального распределения. В этом случае оценкам μ и Σ служат соответственно $\overline{X}$ и S^2, и меры табл. 1.3 могут быть легко вычислены. Коэффициент дивергенции применяется в приложениях дискриминантного (классификационного) анализа [226]. Мера Джеффриса — Матуситы применялась в приложениях дискриминантного анализа к сельскохозяйственным данным для классификации полей [169].

В большинстве работ, указанных в табл. 1.3, рассматриваются одномерные виды мер расстояния. Эти меры обсуждаются в работе Уоккера и Лангриба [383]; там же предлагается их обобщение на многомерный случай. Для более полного ознакомления с мерами, представленными в табл. 1.3, отсылаем читателя к работе [383].

1.6. Кластерные методы, основанные на евклидовой метрике

Основные усилия в развитии методов кластеризации и классификации были направлены на построение методов, основанных на минимизации внутригрупповых сумм квадратов (отклонений). Они могут быть выражены в терминах евклидовых расстояний и называются *методами минимальной дисперсии* [396]. В этом параграфе мы рассмотрим различные методы кластеризации. Кроме

Таблица 1.3. **Многомерные меры расстояния и их метрические свойства**

Название	Вид	Метрика в			Ссылки
		C	MVN	MVN_Σ	
1. Крамер — фон Мизес	$W=\{\int_{-\infty}^{\infty}(G(x)-F(x))^2dx\}^{1/2}$	Да	Да	Да	[67], [381], [76], [304]
2. Колмогоров — Смирнов	$K=\sup_x\lvert G(x)-F(x)\rvert$	Да	Да	Да	[207], [326], [76], [304]
3. Дивергенция	$J=\int_{-\infty}^{\infty}\ln\frac{f(x)}{g(x)}\cdot(f(x)-g(x))dx$	Нет	Нет	Да	[181], [182], [188]
4. Бхаттачария	$B=-\ln\int_{-\infty}^{\infty}[f(x)g(x)]^{1/2}dx$	Нет	Нет	Да	[188], [29]
5. Джеффрис — Матусита	$M=[\int_{-\infty}^{\infty}(\sqrt{g(x)}-\sqrt{f(x)})^2dx]^{1/2}$	Да	Да	Да	[181], [182], [259]
6. Вариационное расстояние Колмогорова	$K(p)=\int_{-\infty}^{\infty}\lvert p_g g(x)-p_f f(x)\rvert dx$	Да	Да	Да	[188], [3], [210]

Продолжение

Название	Вид	Метрика в			Ссылки
		C	MVN	MVN_Σ	
7. Информация Кульбак — Либлер	$L_{fg}=\int_{-\infty}^{\infty}\ln\left(\frac{f(x)}{g(x)}\right)f(x)dx$	Нет	Нет	Да	[216], [188]
8. Свейн — Фу	$T=\frac{\lvert\mu_f-\mu_g\rvert}{D_f+D_g}$, где $D_f=\left\{\frac{\lvert\mu_f-\mu_g\rvert^2(p+2)}{\mathrm{tr}[\Sigma_f(\mu_f-\mu_g)(\mu_f-\mu_g)^T]}\right\}^{1/2}$	Нет	Нет	Да	[353]
9. Махаланобис	$\Delta=\{(\mu_g-\mu_f)^T\Sigma^{-1}(\mu_g-\mu_f)\}^{1/2}$	Не опр.	Не опр.	Да	[251], [252]
10. Самуэль — Бахи	$\mu=\{\int_0^1[F^{-1}(\alpha)-G^{-1}(\alpha)]d\alpha\}^{1/2}$, где $F^{-1}(\alpha)=\inf\{C\mid Q_C\cap Q_\alpha\neq\varnothing\}$ и $Q_C=\{x\mid\sum_{i=1}^{p}x_i\leqslant C\}$, $Q_\alpha=\{x\mid F(x)\geqslant\alpha\}$	Нет	Нет	Нет	[307]
11. Кифер — Вольфовиц	$V=\int_{-\infty}^{\infty}\lvert F(x)-G(x)\rvert e^{-\lvert x\rvert}dx$, где $\lvert x\rvert$ — норма вектора x	Да	Да	Да	[197]

того, мы увидим, что многие приемы кластеризации могут быть охвачены одним алгоритмом посредством общего соотношения, содержащего меры расстояния d_{ij}.

Рассмотрим матрицу наблюдений $X=(X_1, X_2, \ldots, X_n)$. Квадрат евклидова расстояния между X_i и X_j определяется по формуле

$$d_{ij}^2=(X_i-X_j)^T(X_i-X_j).$$

Сейчас мы рассмотрим различные кластерные методы, основанные на этой мере расстояния. Наше описание методов весьма кратко и за деталями отсылаем читателей к соответствующим источникам.

Соренсен [338] описывает так называемый метод полных связей (complete linkage). Суть этого метода заключается в том, что два объекта, принадлежащих одной и той же группе (кластеру), имеют коэффициент сходства, который меньше некоторого порогового значения s. В терминах евклидова расстояния d это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения r. В этом случае r определяет максимально допустимый диаметр подмножества, образующего кластер.

МакНотон-Смит [234] предлагает последовательную процедуру, которую назвал методом максимального локального расстояния (см. определение 1.9); этот метод имеет много общего с предыдущим. Каждый индивид (объект) рассматривается как одноточечный кластер. Объекты группируются последовательно по следующему правилу: два кластера объединяются, если максимальное расстояние между точками одного кластера и точками другого минимально. Процедура состоит из $n-1$ шагов и результатом являются разбиения, которые совпадают со всевозможными разбиениями в методе Соренсена для любых пороговых значений.

Ворд [387] в качестве целевой функции применяет внутригрупповую сумму квадратов (ВСК) отклонений, которая есть не что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. Его метод также представляет собой последовательную процедуру; на каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой

функции, т. е. ВСК. При объединении кластеров I (n_1 элементов) и J (n_2 элементов) это увеличение, как следует из параграфа 1.5, равно:

$$D_{IJ}=\frac{n_1 n_2}{n_1+n_2}(\overline{X}-\overline{Y})^T(\overline{X}-\overline{Y}),$$

где $\overline{X}$ и $\overline{Y}$ обозначают векторы средних по кластерам I и J. Метод Ворда направлен на объединение близко расположенных кластеров.

Сокал и Миченер [334] описывают процедуру, которую назвали *центроидным* методом. Расстояние между двумя кластерами I и J в этом методе определяется как евклидово расстояние между центрами (средними) этих кластеров, т. е. как $d_{IJ}^2=(\overline{X}-\overline{Y})^T(\overline{X}-\overline{Y})$. Кластеризация осуществляется поэтапно [223]: на каждом из $n-1$ шагов объединяют два кластера I и J, имеющие минимальное значение d_{IJ}^2. Если n_1 много больше n_2, то центры $I \cup J$ и I близки друг к другу и характеристики J при объединении кластеров практически игнорируются. Это наталкивает на мысль назвать этот метод методом «взвешенных групп».

Другой метод, предложенный Сокалом и Миченером [334], называется *двухгрупповым* и опирается на связь между объектом i и кластером I. Эта связь выражается в виде среднего коэффициента сходства между объектом i и всеми объектами, входящими в кластер I. Для того чтобы средний коэффициент сходства выразить через евклидово расстояние, обозначим векторы, входящие в кластер I, соответственно через $X_1, X_2, \ldots, X_{n_I}$, а через $\overline{X}$ — центр кластера I. Тогда среднее расстояние D_{iI} между объектом $i \not\subset I$ и всеми объектами из I будет равно:

$$D_{iI}=\frac{1}{n_I}\sum_{j=1}^{n_I}(X_j-Y)^T(X_j-Y),$$

где Y обозначает вектор, соответствующий $i \not\subset I$. Далее

$$D_{iI}=\frac{1}{n_I}\sum_{j=1}^{n_I}(X_j-\overline{X}+\overline{X}-Y)^T(X_j-\overline{X}+\overline{X}-Y)=$$

$$=\frac{1}{n_I}\sum_{j=1}^{n_I}(X_j-\overline{X})^T(X_j-\overline{X})+(\overline{X}-Y)^T(\overline{X}-Y). \qquad (1.11)$$

Первое слагаемое правой части уравнения обозначим S_I^2 и назовем *внутригрупповой дисперсией* объектов из I; второе слагаемое представляет собой квадрат расстояния между объектом i и центром кластера I. Процедура последовательной кластеризации заключается в том, что объект $i \not\subset I$, для которого D_{iI} минимально, присоединяется к кластеру I. Из (1.11) легко видеть, что если два кластера имеют сравнимые дисперсии, то среднее расстояние D_{iI} минимизирует расстояние между объектом i и центром кластера I. Для кластеров с различными дисперсиями объединение происходит в первую очередь с кластером меньшей дисперсии.

Ланс и Уильямс [225] обобщают двухгрупповой метод и определяют среднее сходство между двумя кластерами I и J как среднее сходство между всеми парами объектов из I и J. Этот метод они назвали методом *групповых средних*. Кластеры строятся последовательно; два кластера с минимальным средним коэффициентом сходства объединяются. Для того чтобы среднее сходство выразить в терминах евклидова расстояния, обозначим через $\overline{X}$ и $\overline{Y}$ соответственно средние кластеров I и J. Средний квадрат расстояний между объектами кластеров I и J, обозначенный через D_{IJ}^2, будет равен

$$D_{IJ}^2 = \frac{1}{n_I n_J} \sum_{i=1}^{n_I} \sum_{j=1}^{n_J} (X_i - Y_j)^T (X_i - Y_j) =$$

$$= \frac{1}{n_I} \sum_{i=1}^{n_I} \left\{ \frac{1}{n_J} \sum_{j=1}^{n_J} (Y_j - \overline{Y})^T (Y_j - \overline{Y}) + \right.$$

$$\left. + (\overline{Y} - X_i)^T (\overline{Y} - X_i) \right\} = \frac{1}{n_J} \sum_{j=1}^{n_J} (Y_j - \overline{Y})^T (Y_j - \overline{Y}) +$$

$$+ \frac{1}{n_I} \sum_{i=1}^{n_I} (\overline{Y} - X_i)^T (\overline{Y} - X_i).$$

Первое слагаемое правой части выражения есть внутригрупповая дисперсия кластера J, а второе слагаемое — средний квадрат расстояний между X_i и $\overline{Y}$, $i = 1, 2, \ldots, n_I$. Таким образом, второе слагаемое может быть переписано как

$$S_I^2 + d^2(\overline{X}, \overline{Y}) = \frac{1}{n_I} \sum_{i=1}^{n_I} (X_i - \overline{X})^T (X_i - \overline{X}) +$$

$$+ (\overline{Y} - \overline{X})^T (\overline{Y} - \overline{X}),$$

откуда

$$D_{IJ}^2 = S_I^2 + S_J^2 + d^2(\overline{X}, \overline{Y}), \qquad (1.12)$$

т. е. минимизация среднего сходства эквивалентна минимизации (1.12).

Боннер [33] описывает метод, в котором объект, служащий начальной точкой, выбирается случайно. Все объекты, лежащие на расстоянии от начальной точки не больше r, принимаются за первый кластер. Из оставшихся точек снова случайным образом выбирается объект и процесс повторяется, как и предшествовавший. В результате все точки будут разбиты на группы.

Хиверинен [171] рассматривает процедуру, аналогичную Боннеру, но в качестве начального объекта кластеризации выбирает не случайную, а так называемую «типическую» точку. Для определения «типических» точек он пользуется статистикой потери информации, причем эти точки лежат на минимальном расстоянии от центра оставшегося множества объектов.

В процедуре Болла и Холла [18] первоначальные K кластеров формируются случайным отбором K точек, к которым затем присоединяется каждая из оставшихся $n - K$ точек — по минимальному расстоянию к той или иной из них. Затем находятся центры кластеров и два кластера I и J объединяются, если D_{IJ}^2 меньше некоторого порогового значения r. Наоборот, если внутригрупповая дисперсия кластера S_x^2 по некоторой переменной x превосходит пороговое значение S^2, то кластер разбивается. Таким образом, дисперсии S_I^2 кластеров, получающихся в результате этой процедуры, ограничены:

$$S_I^2 \leqslant pS^2,$$

где p — число переменных. Вместо центра первоначального кластера рассматриваются центры новых образо-

вавшихся кластеров и процесс продолжается до тех пор пока не сойдется. Процедура Болла и Холла становится довольно популярной.

МакКвин [237] предлагает метод, аналогичный методу Болла и Холла. Случайным образом отбирается k объектов, которые принимаются в качестве центров кластеризации. Для каждого объекта отыскивается ближайшая точка кластеризации, и если расстояние от выбранного объекта до этой точки не больше заданного уровня r, то объект приписывают к кластеру найденной точки кластеризации. Если это расстояние больше r, то объект образует новый кластер. После этого вычисляются новые центры кластеров. Если расстояние между центрами двух кластеров меньше другого априорно заданного уровня, то соответствующие кластеры объединяются. Процесс продолжается до сходимости.

Метод Себестьена [315] имеет много общего с предыдущим. Однако по Себестьену объект принадлежит кластеру, если расстояние d до центра кластера меньше r; если же это расстояние больше R $(R>r)$, то этот объект образует новую точку кластеризации. Однако если $r<d<R$, то объект выбывает из рассмотрения до следующей итерации.

Дженси [174] предложил процедуру, сходную с предложенной МакКвином. Однако в методе Дженси случайным образом выбирается k точек не из n рассматриваемых объектов, как в методе МакКвина, а из всего пространства E_p. В качестве минимизируемой целевой функции берется внутригрупповая сумма квадратов отклонений.

Форджи [114] рассматривает метод, сходный с методом Дженси. Разбиение объектов на кластеры в этом методе близко к разбиению, предложенному Дженси. Здесь также пользуются минимизацией внутригрупповой суммы квадратов.

Основная причина популярности евклидовой метрики в кластерном анализе заключается скорее всего в том, что она наиболее близка к интуитивному представлению о расстоянии, а также, как следует из уравнения (1.7), в том, что она тесно связана с ВСК.

Имеются также и возражения против подхода, основанного на минимальной дисперсии в кластерном анализе. Так, изменение масштаба приведет к другому разбиению на кластеры. С этими и другими возражениями

и их обсуждением читатель может ознакомиться по работе Уишарта [396]. Там же рассматриваются методы, описанные выше. Фридман и Рубин [122] обсуждают некоторые инвариантные критерии группировки наблюдений.

1.7. Алгоритм последовательной кластеризации

Схема последовательной кластеризации может быть описана следующим образом. Рассмотрим $I=\{I_1, I_2, \ldots, I_n\}$ как множество кластеров $\{I_1\}, \{I_2\}, \ldots, \{I_n\}$; выберем два из них, скажем I_i и I_j, которые в некотором смысле наиболее близки друг к другу, и объединим их в один кластер. Новое множество кластеров, состоящее уже из $n-1$ кластеров, будет

$$\{I_1\}, \{I_2\}, \ldots, \{I_i, I_j\}, \ldots, \{I_n\}.$$

Повторяя процесс, мы получим последовательные множества кластеров, состоящие из n—2, n—3 и т. д. кластеров. В конце этой процедуры получится кластер, состоящий из n объектов и совпадающий с первоначальным множеством $I=\{I_1, I_2, \ldots, I_n\}$.

В качестве меры расстояния примем квадрат евклидовой метрики d_{ij}^2. Для наглядности вычислим матрицу $D=\{d_{ij}^2\}$, где d_{ij}^2 — квадрат расстояния между I_i и I_j.

Таблица 1.4. Значения d_{ij}^2

	I_1	I_2	I_3	. . .	I_n
I_1	0	d_{12}^2	d_{13}^2	. . .	d_{1n}^2
I_2		0	d_{23}^2	. . .	d_{2n}^2
I_3			0	. . .	d_{3n}^2
.			.		
.				.	
.					
I_n					0

Предположим, что расстояние между I_i и I_j минимально, т. е. что $d_{ij}^2=\min\{d_{ij}^2,\ i\neq j\}$; образуем с помощью I_i и I_j новый кластер $\{I_i, I_j\}$. Построим новую $(n-1)\times(n-1)$ матрицу расстояния (см. табл. 1.5).

Таблица 1.5. **Значения d_{ij}^2 после первого объединения**

	$\{I_i, I_j\}$	I_1	I_2	I_3	...	I_n
$\{I_i, I_j\}$	0	d_{ij1}^2	d_{ij2}^2	d_{ij3}^2	...	d_{ijn}^2
I_1		0	d_{12}^2	d_{13}^2	...	d_{1n}^2
I_2			0	d_{23}^2	...	d_{2n}^2
I_3				0	...	d_{3n}^2
.					.	
.					.	
.					.	
I_n						0

Легко видеть, что $n-2$ строки для этой матрицы можно непосредственно взять из старой, однако первую строку необходимо вычислить заново. Очевидно, вычисления будут сведены к минимуму, если удастся выразить d_{ijk}^2, $k=1, 2, \ldots, n$, $k \neq i \neq j$ через элементы первоначальной матрицы.

Ланс и Уильямс [223] предложили рекурсивную процедуру, в которой вычисления матрицы расстояний опираются только на значения расстояний в предыдущей матрице. Их рекурсивная схема предполагает использование минимального и максимального локальных расстояний, медианы, групповых средних и центра. Все пять случаев, за исключением минимального локального расстояния и медианы, были описаны в параграфе 1.6. Минимальное локальное расстояние между двумя кластерами I и J было определено в параграфе 1.5; схема предусматривает объединение двух кластеров, имеющих наименьшее минимальное локальное расстояние. Медианный метод такой же, как и центроидный, за исключением того, что здесь при объединении кластеров I и J предполагается, что $n_I = n_J$, и поэтому центр нового кластера лежит точно посередине между центрами старых кластеров.

Уишарт [394] считает, что процедуру Уорда [387] можно объединить с пятью, рассмотренными выше. Как было показано в параграфе 1.5, объединение кластеров

I и J ведет к увеличению ВСК на величину W_{IJ}, которая задается равенством

$$W_{IJ} = \frac{n_I n_J}{n_I + n_J} (\overline{X}_I - \overline{Y}_J)^T (\overline{X}_I - \overline{X}_J) = \frac{n_I n_J}{n_I + n_J} d_{IJ}^2, \quad (1.13)$$

где $d_{IJ}^2 = (\overline{X}_I - \overline{X}_J)^T (\overline{X}_I - \overline{X}_J)$. Если кластер $I \cup J = L$ объединяется с K, то можно показать, что $d_{KL}^2 = (\overline{X}_K - \overline{X}_L)^T (\overline{X}_K - \overline{X}_L)$ и

$$d_{KL}^2 = \frac{n_I}{n_L} d_{KI}^2 + \frac{n_J}{n_L} d_{KJ}^2 - \frac{n_I n_J}{n_L^2} d_{IJ}^2 . \quad (1.14)$$

Более того, из (1.13) следует, что

$$W_{KL} = \frac{n_K n_L}{n_K + n_L} d_{KL}^2 . \quad (1.15)$$

Подставляя выражение (1.15) для каждого d^2 в уравнение (1.14), получим:

$$W_{KL} = \frac{1}{n_K + n_L} \{(n_I + n_K) W_{KI} + (n_J + n_K) W_{KJ} - n_K W_{IJ}\}. \quad (1.16)$$

Уравнение (1.16) определяет величину приращения ВСК при объединении K и $I \cup J$.

Начиная с матрицы квадратов евклидовых расстояний (табл. 1.4) $D = \{d_{ij}^2, i=1, 2, \ldots, n; j=1, 2, \ldots, n\}$ процедура заключается в объединении таких кластеров I_p и I_q, для которых $d_{pq}^2 = 2W_{pq}$ минимально. Кластер I_p, состоящий из одного объекта, заменяется на $I_p \cup I_q$, а расстояния d_{ip}^2, $i=1, 2, \ldots, n$; $i \neq p$, q в матрице D заменяются на $d_{ip}^2 = 2W_{ip}$. Элементы q-го столбца и строки полагаются равными нулю, т. е. S_q становится «недействительным» [394]. Соответственно n_p заменяется на $n_p + n_q$, а n_q приравнивается нулю. Равенство

$$d_{ip}^2 = 2W_{ip} \quad (1.17)$$

выполняется для всех $\{d_{ij}^2\}$, $i, j \neq q$.

Подставляя W_{ip} (фактически W_{ir}) из уравнения (1.16) в уравнение (1.17), получим

$$d_{ip}^2 = 2W_{ir} = \frac{2}{(n_i + n_r)} [(n_i + n_p) W_{ip} +$$

$$+(n_i+n_q)W_{iq}-n_iW_{pq}]=\frac{1}{(n_i+n_r)}[(n_i+$$

$$+n_p)d_{ip}^2+(n_i+n_q)d_{iq}^2-n_id_{pq}^2], \qquad (1.18)$$

где $n_r=n_p+n_q$. Если на каждом шаге объединения p-е столбцы и строки матрицы D преобразовываются по формуле (1.18), то равенство (1.17) будет выполняться для всех d_{ij}^2 и всех действительных множеств S_i и S_j. Заметим, что d_{ij}^2 в (1.17) не является евклидовым расстоянием, если не рассматриваются только два кластера, содержащих по одному элементу.

Алгоритм группировки окончательно может быть записан следующим образом [394]:

1) Найдем $d_{pq}^2=\min\{d_{ij}^2\}$, $i=1,\ldots,\ j-1$; $j=2,\ldots,n$; $n_i>0$; $n_j>0$;

2) Увеличение целевой функции при объединении двух кластеров I_p I_q равно $W_{pq}=\frac{1}{2}d_{pq}^2$;

3) I_p заменяется на $S_p\cup S_q$; строка $\{d_{ip}^2\}$ и столбец $\{d_{pj}^2\}$ матрицы D пересчитываются по формуле (1.18), $i=1,\ 2,\ldots,\ p-1$; $n_i>0$; $j=p+1,\ldots,n$; $j\neq q$; $n_j>0$;

4) Полагаем $n_p=n_p+n_q$ и $n_q=0$, превращая S_q в недействительное множество;

5) Записываем элементы кластера S_q в кластер S_p, возвращаемся к (1) и повторяем процедуру $n-2$ раз.

Ланс и Уильямс [223] получили общее уравнение (1.19), аналогичное (1.18) и верное для всех пяти процессов кластеризации, описанных выше. Это уравнение может быть записано в следующем виде:

$$d_{hk}^2=\alpha_id_{hi}^2+\alpha_jd_{hj}^2+\beta d_{ij}^2+\gamma|d_{hi}^2-d_{hj}^2|, \qquad (1.19)$$

где α_i, α_j, β и γ — параметры, задающие вид процесса. В уравнении (1.19) d_{hk} обозначает меру расстояния между кластерами I_h и $I_k=I_i\cup I_j$.

Значения параметров, входящих в общую формулу (1.19), соответствующие шести различным процессам кластеризации, приведены ниже:

минимальное локальное расстояние: $\alpha_i=\alpha_j=\frac{1}{2}$; $\beta=0$; $\gamma=-\frac{1}{2}$;

максимальное локальное расстояние: $\alpha_i=\alpha_j=\frac{1}{2}$; $\beta=0$; $\gamma=\frac{1}{2}$;

медиана: $\alpha_i=\alpha_j=\frac{1}{2}$; $\beta=-\frac{1}{4}$; $\gamma=0$;

среднее группы: $\alpha_i=n_i/n_k$; $\alpha_j=n_j/n_k$; $\beta=\gamma=0$;

центроидный метод: $\alpha_i=n_i/n_k$; $\alpha_j=n_j/n_k$; $\beta=-\alpha_i\alpha_j$; $\gamma=0$;

метод Уорда: $\alpha_i=\frac{n_h+n_i}{n_h+n_k}$; $\alpha_j=\frac{n_h+n_j}{n_h+n_k}$; $\beta=\frac{-n_h}{n_h+n_h}$; $\gamma=0$.

Первые пять значений параметров приводятся в работе Ланса и Уильямса [223]. Значения параметров в методе Уорда найдены Уишартом [394] и получаются из уравнения (1.19). Все шесть методов были объединены в одну вычислительную программу с параметрами α, β и γ [398], [399].

1.8. Другие вопросы кластерного анализа

Одним из важнейших вопросов при решении кластерной проблемы является выбор необходимого числа кластеров. В некоторых случаях число кластеров m может быть выбрано априорно, однако в общем случае это число определяется в процессе разбиения множества на кластеры. В этой книге мы не будем подробно останавливаться на этой сложной проблеме.

Хорошо известно, что в некоторых задачах с большим числом наблюдений для практических целей пользуются методом случайного отбора. Фортьер и Соломон исследовали эти методы [119] и нашли, что законы простого случайного отбора могут быть применены для вычисления числа кластеров, которое должно быть принято для достижения вероятности α того, что найдено наилучшее разбиение. Таким образом, оптимальное число разбиений является функцией заданной доли β «наилучших» или в некотором смысле допустимых разбиений в множестве всех возможных. Общее рассеяние множества кластеров будет тем больше, чем выше доля β «допустимых» разбиений. Фортьер и Соломон приводят таблицу, по которой можно найти необходимое число разбиений $S(\alpha, \beta)$ в зависимости от значений α и β. При этом в качестве меры разнородности рассматрива-

ется не мера рассеяния, а «мера принадлежности», введенная Хользингером и Харманом [168] (см. табл. 1.6). Фортьер и Соломон пришли к выводу, что простой случайный отбор в общем случае не эффективен, если распределение показателя очень скошено и более вероятные его значения находятся на хвостах распределения. В то же время, как отмечают авторы, «модификация стратегии отбора может значительно улучшить ситуацию и эту возможность необходимо исследовать».

Таблица 1.6. Значения S (α, β)

β \ α	0,20	0,10	0,05	0,01	0,001	0,0001
0,20	8	11	14	21	31	42
0,10	16	22	29	44	66	88
0,05	32	45	59	90	135	180
0,01	161	230	299	459	689	918
0,001	1626	2326	3026	4652	6977	9303
0,0001	17475	25000	32526	55000	75000	100000

При решении задачи кластерного анализа молчаливо принимается, что 1) выбранные характеристики в принципе допускают желательное разбиение на кластеры, 2) единицы измерения (масштаб) выбраны правильно. Первая проблема называется проблемой выбора свойств или характеристик объектов; этому вопросу посвящены работы [229], [230] и [255]. Вообще предполагается, что проблема выбора характеристик решена до начала процесса кластеризации. Однако следует предупредить, что этим вносится некоторый произвол, что в отдельных случаях требует дополнительного рассмотрения.

Другой вопрос, который всегда сопутствует измерению, — выбор масштаба — также играет большую роль. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклонение; так что дисперсия оказывается равной единице. В случае же, когда исходят из непосредственных (обычных) единиц измерения, возникает проблема интерпретации. Однако наиболее серьезная проблема возникает в связи с тем, что разбиение на кластеры зависит от выбора масштаба. Было бы желательно иметь такой метод кластеризации, который был бы инвариантен к изменению масштабов измерения.

ГЛАВА 2

КЛАСТЕРИЗАЦИЯ ПОЛНЫМ ПЕРЕБОРОМ

2.1. Введение

Наиболее прямой способ решения кластерной проблемы заключается в полном переборе всех возможных разбиений на кластеры и отыскании такого разбиения, которое ведет к оптимальному (минимальному) значению целевой функции. Однако такая процедура практически невыполнима за исключением тех случаев, когда n (число объектов) и m (число кластеров) *не велико*. Этот способ называют *кластеризацией с помощью полного перебора*. Например, если $n=8$, а $m=4$, то число возможных разбиений равно 1701; другими словами, существует 1701 способов разбить 8 объектов на 4 группы. Число разбиений обозначается через $S(n, m)$ и называется числом Стирлинга второго рода. Оно может быть вычислено по формулам, которые будут нами получены в следующем параграфе. Данная глава лишь частично имеет отношение к проблеме кластеризации и поэтому при первом чтении может быть опущена.

2.2. Число разбиений n объектов на m непустых подмножеств

Процесс разбиения множества из n объектов на m непустых подмножеств можно представить как распределение n различных шаров по m одинаковым урнам, ни одна из которых не должна остаться пустой.

Обозначим число таких разбиений через w. Тогда $wm!$ есть число всех возможных размещений n различ-

ных объектов по m разным урнам (ни одна из которых не остается пустой)*.

Один из наиболее эффективных методов решения задач, связанных с этой проблемой (т. е. разбиение n объектов на m непустых подмножеств), основан на методе производящих функций.

Рассмотрим функцию

$$(1+x_1t)(1+x_2t)\ldots(1+x_nt). \qquad (2.1)$$

Раскрывая скобки, мы получим полином от t, в котором коэффициент при t^k представляет собой сумму произведений всех комбинаций по k множителей из n.

Если, далее, $x_1=x_2=\ldots=x_n=1$, то коэффициент при t^k равен C_n^k — числу способов отбора (сочетаний) k объектов из n. Таким образом, последовательность C_n^k можно получить при помощи *производящей функции*

$$(1+t)^n=\sum_{k=0}^{n} C_n^k t^k .$$

Этой же функцией можно воспользоваться в случае, когда порядок элементов в группе существен:

$$(1+t)^n=\sum_{k=0}^{n} C_n^k t^k=\sum_{k=0}^{n} A_n^k \frac{t^k}{k!}, \qquad (2.2)$$

здесь $A_n^k=k!\,C_n^k$, $k=0, 1, \ldots, n$ (число размещений). Функция (2.2) называется *экспоненциальной производящей функцией*. Если допускаются повторения, то для подсчета результата перебора служит ряд, k-й член которого равен $t^k/k!$, где k — число повторений. Таким образом, если число повторений не ограничено, то счетчиком для каждого объекта служит ряд

$$1+t+\frac{t^2}{2!}+\ldots=e^t .$$

Следовательно, число различных перестановок n различных объектов в группе из k объектов равно коэффициенту при $t^k/k!$ в производящей функции

$$\left(1+t+\frac{t^2}{2!}+\ldots\right)^n=e^{nt}=\sum_{k=0}^{\infty} n^k\frac{t^k}{k!} .$$

Как видим, это число равно n^k. Если число повторений не ограничено и каждый объект повторяется не менее одного раза, то счетчиком будет

$$\left(t+\frac{t^2}{2!}+\ldots\right)^n=(e^t-1)^n=\sum_{j=0}^{n} C_n^j(-1)^j e^{(n-j)t}=$$
$$=\sum_{k=0}^{\infty}\frac{t^k}{k!}\left(\sum_{j=0}^{n} C_n^j(-1)^j(n-j)^k\right). \qquad (2.3)$$

Тогда число перестановок определяется формулой

$$\sum_{j=0}^{n} C_n^j(-1)^j(n-j)^k . \qquad (2.4)$$

Производящая функция, применяемая для получения числа способов распределения n различных шаров по m различным урнам (порядок расположения шаров в урнах несуществен), при котором урна i содержит n_i шаров, $i=1, 2, \ldots, m$ задается как

$$(x_1+x_2+\ldots+x_m)^n=\sum\frac{n!}{n_1!\,n_2!\ldots n_m!}x_1^{n_1}x_2^{n_2}\ldots x_m^{n_m}, \qquad (2.5)$$

где суммирование производится по всем m-разбиениям n (так что $n_1+n_2+\ldots+n_m=n$). Коэффициент при $x_1^{n_1}x_2^{n_2}\ldots x_m^{n_m}$ равен числу таких распределений n шаров, при которых урна 1 содержит n_1 шаров, урна 2 содержит $n_2, \ldots$, урна m содержит n_m шаров. Таким образом, величина

$$\frac{n!}{n_1!\,n_2!\ldots n_m!} \qquad (2.6)$$

есть число способов размещения n_i шаров в i-й урне, $i=1, 2, \ldots, m$.

Число способов размещения n различных шаров по m различным урнам может быть найдено с помощью производящей функции

$$\prod_{i=1}^{m}\left(1+x_it+x_i^2\frac{t^2}{2!}+\ldots+x_i^n\frac{t^n}{n!}\right). \qquad (2.7)$$

В этом случае i-й множитель

---

* Число w не предполагает различения и самих урн. Если же учитывать и его, то, расположив для данного разбиения урны в некотором порядке, будем их переставлять, обменивая содержащиеся в них группы объектов. Таких перестановок может быть $m!$

Обозначения понятий теории соединений дальше изменены на привычные для русского читателя.— *Примеч. ред.*

$$1+x_i t+x_i^2\frac{t^2}{2!}+\ldots+x_i^n\frac{t^n}{n!} \qquad (2.8)$$

соответствует *числу способов размещения шаров* в i-й урне. Раскрывая скобки в (2.7), находим, что коэффициент при t^n равен

$$\sum\frac{1}{n_1!n_2!\ldots n_m!}x_1^{n_1}x_2^{n_2}\cdots x_m^{n_m},$$

где суммирование производится по всем m-разбиениям n. Таким образом, коэффициент при $t^n/n!$ в (2.7) равен:

$$\sum\frac{n!}{n_1!\,n_2!\ldots n_m!}x_1^{n_1}x_2^{n_2}\cdots x_m^{n_m}=$$
$$=(x_1+x_2+\cdots+x_m)^n.$$

Как следует из (2.6), число способов размещения n шаров по m урнам, при которых k-я урна содержит n_k шаров ($k=1,2,\ldots,m$), равно $n!/n_1!\,n_2!\ldots n_m!$, так как $x_1^{n_1}x_2^{n_2}\ldots x_m^{n_m}$ отвечает некоторому частному способу размещения. Полагая в (2.7) $x_1=x_2=\ldots=x_m=1$, получаем, что коэффициент при $t^n/n!$ в выражении

$$(1+t+t^2/2!+\cdots+t^n/n!)^m$$

есть число способов размещения n различных шаров по m различным урнам. Однако этот коэффициент равен коэффициенту при $t^n/n!$ в выражении

$$(1+t+t^2/2!+\cdots)^m=e^{mt}.$$

Отсюда следует, что число возможных размещений n шаров по m урнам равно m^n*.

Число разбиений множества из n объектов на m подмножеств, ни одно из которых не будет пустым, определяется в следующей теореме.

Теорема 1.1. Число способов размещения n шаров по m урнам (ни одна из урн не будет пустой, а порядок расположения шаров в урне несуществен) равно

$$\sum_{j=0}^{m}C_m^j(-1)^j(m-j)^n.$$

* Допускается, что та или иная урна может вообще не содержать шаров, а поэтому этот же результат может быть получен крайне просто: первый шар можно положить в любую из m урн. Так же второй и т. д. до n-го. Отсюда сразу ясно, что число всех полученных вариантов в целом будет m^n. — *Примеч. ред.*

Доказательство. Число шаров, содержащихся в i-й урне, определяется из уравнения

$$(x_i t+x_i^2 t^2/2!+\cdots)=e^{tx_i}-1,$$

поскольку каждая урна содержит по крайней мере один шар. Соответствующая экспоненциальная производящая функция, которую обозначим через $E(t,x_1,,,,,x_m)$, равна:

$$E(t,x_1,x_2,\ldots,x_m)=\prod_{i=1}^{m}(e^{tx_i}-1). \qquad (2.9)$$

Число способов размещения n шаров по m урнам так, чтобы ни одна урна не была пустой, равно коэффициенту при $t^n/n!$ и при условии $x_1=x_2=\ldots=x_m=1$. В этом случае производящая функция (2.9) превращается в

$$E(t,1,1,\ldots,1)=(e^t-1)^m, \qquad (2.10)$$

которая совпадает с (2.3). Поэтому

$$E(t,1,1,\ldots,1)=\sum_{k=0}^{\infty}\frac{t^k}{k!}\Big[\sum_{j=0}^{m}C_m^j(-1)^j(m-j)^k\Big].$$

Коэффициент при $t^n/n!$ равен

$$\sum_{j=0}^{m}C_m^j(-1)^j(m-j)^n,$$

что приводит к требуемому результату.

В теореме 1.1 m урн предполагаются различными. Однако при разбиении n объектов на m подмножеств, каждое из которых не пусто, порядок подмножеств не играет роли. С учетом теоремы 1.1 отсюда следует, что число разбиений n объектов на m подмножеств равно:

$$w=\frac{1}{m!}\sum_{j=0}^{m}C_m^j(-1)^j(m-j)^n. \qquad (2.11)$$

Осталось показать, каким образом числа Стирлинга второго рода связаны с соотношением (2.11).

Числа Стирлинга второго рода возникают при вычислении конечных разностей и связаны с нахождением показателя x в *факториальных многочленах*. Факториальный многочлен определяется следующим образом:

$$x_{(0)}=1,\ x_{(n)}=x(x-1)\cdots(x-n+1),\ n=1,2,\cdots$$

Определение 2.1. Числами Стирлинга второго рода называются числа $S(n, i)$, удовлетворяющие уравнениям

$$x^n=\sum_{i=1}^{n} S(n,i)x_{(i)},\ S(n,0)=0$$

и

$$S(n,n+k)=0 \text{ для } k>0.$$

По теореме Ньютона [172, гл. 6] полином n-й степени $U(x)$ может быть записан в виде:

$$U(x)=\sum_{k=0}^{n} x_{(k)}\left[\frac{\Delta^k U(x)}{k!}\right]_{x=0}, \qquad (2.12)$$

где Δ определяется как

$$\Delta f(x)=f(x+1)-f(x),$$

$$\Delta^k f(x)=\Delta[\Delta^{k-1}f(x)]$$

и

$$\Delta^0 f(x)=f(x).$$

Разлагая $U(x)=x^n$ по формуле (2.12), получим:

$$x^n=\sum_{m=0}^{n} x_{(m)}\left[\frac{\Delta^m x^n}{m!}\right]_{x=0}.$$

Из определения $S(n, i)$ имеем:

$$S(n,m)=\left[\frac{\Delta^m x^n}{m!}\right]_{x=0}=\frac{1}{m!}[\Delta^m x^n]_{x=0}.$$

Осталось показать, что

$$\sum_{j=0}^{m} C_m^j(-1)^j(m-j)^n=[\Delta^m x^n]_{x=0}.$$

Для этой цели введем оператор смещения $E=1+\Delta$, где $Ef(x)=f(x+1)$, $E^j f(x)=f(x+j)$ и $\Delta=E-1$. Пользуясь введенными обозначениями, получим:

$$\sum_{k=0}^{m} C_m^k(-1)^k(m-k)^n=\sum_{k=0}^{m} C_m^k(-1)^k E^{m-k}x^n\big|_{x=0}=$$

$$=(E-1)^m x^n\big|_{x=0}=[\Delta^m x^n]_{x=0}.$$

Таким образом, имеем $w=S(n, m)$.

Если число подмножеств разбиения m известно заранее, то общее число разбиений n объектов на m кластеров равно $S(n, m)$. Если же значение m неизвестно, то общее число всех разбиений равно

$$\sum_{m=1}^{n} S(n,m).$$

2.3. Рекурсивное соотношение между числами Стирлинга второго рода

Для нахождения чисел Стирлинга второго рода может быть применено уравнение (2.1). Однако если требуется найти последовательность значений этих чисел, то удобнее воспользоваться рекуррентным соотношением, связывающим эти величины

$$S(n+1,i)=iS(n,i)+S(n,i-1).$$

Эта формула следует прежде всего из того, что

$$x_{(i+1)}=x(x-1)\cdots(x-i)=x_{(i)}(x-i)=xx_{(i)}-ix_{(i)}.$$

Откуда

$$x_{(i+1)}+ix_{(i)}=xx_{(i)}. \qquad (2.13)$$

Пользуясь определением $S(n, i)$, найдем:

$$x^{n+1}=\sum_{i=1}^{n+1} S(n+1,i)x_{(i)} \qquad (2.14)$$

и

$$x^{n+1}=xx^n=\sum_{i=1}^{n} S(n,i)xx_{(i)}. \qquad (2.15)$$

Воспользовавшись равенством (2.13), из (2.15) получим:

$$x^{n+1}=\sum_{i=1}^{n} S(n,i)\,(x_{(i+1)}+ix_{(i)})=$$

$$=\sum_{i=2}^{n+1} S(n,i-1)x_{(i)}+\sum_{i=1}^{n} iS(n,i)x_{(i)}.$$

Но поскольку $S(n, 0)=S(n, n+1)=0$,

$$x^{n+1}=\sum_{i=1}^{n+1}[S(n,i-1)+iS(n,i)]x_{(i)}. \qquad (2.16)$$

Сравнивая этот результат с (2.4), мы имеем:

$$S(n+1,i)=S(n,i-1)+iS(n,i).$$

Эквивалентной формулой для $S(n, m)$ является

$$S(n, m)=\frac{1}{m!}\sum_{k=0}^{m} C_m^k (-1)^{m-k} k^n ,$$

откуда получаем:

$$\frac{S(n, m)}{m^n}=\frac{1}{m!}\sum_{k=0}^{m} C_m^k (-1)^{m-k} \left(\frac{k}{m}\right)^n .$$

При $n\to\infty$ каждый член суммы за исключением последнего стремится к нулю, поэтому

$$\lim_{n\to\infty} \frac{S(n, m)}{m^n}=\frac{1}{m!} .$$

Таким образом, при больших n

$$S(n, m)\approx\frac{m^n}{m!}=m^{n-1} .$$

В табл. 2.1 приводятся значения $S(n, m)$ для n, не превышающих 8.

Таблица 2.1. **Число разбиений на кластеры для различных значений m и n**[1]

n \ m	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	1	3	1					
4	1	7	6	1				
5	1	15	25	10	1			
6	1	31	90	65	15	1		
7	1	63	301	350	140	21	1	
8	1	127	966	1701	1050	266	28	1

[1] n — число объектов, m — число кластеров.

2.4. Вычислительные аспекты полного перебора

Имея целевую функцию, например внутригрупповую сумму квадратов, которая эквивалентна евклидовой метрике, оптимальное решение кластерной проблемы теоретически может быть получено перебором всех возможных альтернатив разбиения; при этом оптимальному разбиению отвечает и оптимальное значение целевой функции. Однако этот процесс практически неосуществим за исключением случаев, когда n мало. При полном переборе некоторые кластеры для разных альтернатив разбиения будут совпадать, что приводит к лишним вычислениям, которые можно сократить.

Поэтому при кластеризации методом полного перебора желательно иметь такую схему или алгоритм, которые исключали бы излишнюю вычислительную работу. Это приводит к рассмотрению методов динамического программирования, которые, значительно сокращая общее число вычислений, в то же время сходятся к оптимальному решению. Рассмотренные в первой главе методы кластеризации также достаточно быстро приводят к оптимальному решению. К таким методам, например, принадлежат методы последовательной кластеризации. Однако эти методы работают только на подмножестве всех возможных альтернатив разбиения и поэтому не гарантируют того, что найденное решение будет оптимальным или близким к нему.

При кластеризации с помощью полного перебора необходимо иметь в памяти одну матрицу наблюдений X, и все вычисления будут основываться только на ней без привлечения дополнительных массивов. Однако необходимый объем вычислений так велик, что, несмотря на высокое быстродействие современных вычислительных машин, решение задачи остается безнадежным. (В табл. 2.1 приводится общее число кластеров для различных значений n (число объектов меньших 9).)

Привлечение методов динамического программирования для решения кластерной проблемы требует большой скорости при обращении к дополнительным массивам, хранящим информацию. Эти методы привлекают «старые» вычисления и выполняют «новые». Таким образом, высокая скорость обращения к хранящейся информации очень желательна, если не обязательна. Один из методов динамического программирования был разработан Дженсеном [183]; этот метод будет описан в следующей главе. Там же будут рассмотрены другие методы, основанные на динамическом программировании.

ГЛАВА 3

МАТЕМАТИЧЕСКОЕ ПРОГРАММИРОВАНИЕ И КЛАСТЕРНЫЙ АНАЛИЗ

Напомним, что решением кластерной задачи является такое разбиение n объектов на m непересекающихся подмножеств, которое удовлетворяет некоторому критерию однородности внутри кластеров. Один из способов отыскания такого разбиения был рассмотрен в предыдущей главе: он заключался в полном переборе всех возможных разбиений на m кластеров и выбора из них оптимального. К сожалению, указанный метод практически неосуществим даже для небольших n и m.

В качестве альтернативы метода полного перебора можно предложить другие методы, составляющие содержание так называемого математического программирования; эти методы позволяют сократить общий объем вычислений и в то же время приводят к оптимальному решению. Заметим, что большинство из ранее рассмотренных методов кластеризации (гл. 1) дает оптимальное решение в классе меньшем, чем класс всех возможных разбиений (кластеров), поэтому нет гарантии, что найденное решение будет оптимально в классе всех разбиений. Различные применения математического программирования читатель найдет, например, у Дженсена [183], Вайнода [380] и Рао [291].

3.1. Применение динамического программирования к кластерному анализу

В этом параграфе мы рассмотрим задачу разбиения множества из 6 объектов на 3 подмножества; в качестве меры расстояния между объектами выберем евклидову метрику, что соответствует критерию минимизации внутригрупповой суммы квадратов (ВСК).

Напомним, что ВСК вычисляется по формуле

$$W=\operatorname{tr}\sum_{j=1}^{m}S_j=\sum_{j=1}^{m}W_j\,,$$

где S_j обозначает $p\times p$ матрицу рассеяния j-го кластера, а $\operatorname{tr}S_j=W_j$. Таким образом, имеем:

$$W=\sum_{l=1}^{m}\left(\frac{1}{2n_l}\sum_{i=1}^{n_l}\sum_{j=1}^{n_l}d^2(X_i,X_j)\right)=$$

$$=\sum_{l=1}^{m}\left(\frac{1}{2n_l}\sum_{i=1}^{n_l}\sum_{j=1}^{n_l}d_{ij}^{\,2}\right)\,, \qquad (3.1)$$

где $d^2(X_i, X_j)=(X_i-X_j)^T(X_i-X_j)$.

Суть методов динамического программирования состоит в целенаправленном поиске разбиения, дающего минимальное значение величины W, при этом разбиения, которые приводят к большему значению W, отбрасываются.

Подробнее остановимся на проблеме разбиения $n=6$ объектов на $m=3$ группы с помощью полного перебора. На этом же примере мы рассмотрим применения методов динамического программирования, которые приведены в работе Дженсена [183].

Общее число способов разбиения 6 объектов на 3 группы определяется из уравнения (2.11):

$$S(6,3)=\frac{1}{3!}\sum_{k=0}^{3}(-1)^kC_3^k\,(3-k)^6=90.$$

90 альтернатив кластеризации может быть классифицировано соответственно *формам распределения* [183]. В нашем примере существуют три формы распределения, которые обозначим

(1) {4} {1} {1};
(2) {3} {2} {1};
(3) {2} {2} {2}.

Каждая компонента формы распределения обозначает число объектов в некотором кластере. Компоненты в форме распределения всегда будем записывать в убывающем порядке. В нашем примере имеется 90 альтернатив разбиения и 3 формы распределения. В общем

случае число форм распределения значительно меньше, чем число возможных разбиений.

Всего существует $C_6^4 C_2^1/2=15$ видов разбиения, соответствующих первой форме распределения {4}, {1}, {1}; $C_6^3 C_3^2=60$ видов разбиения, соответствующих второй форме распределения {3}, {2}, {1} и $C_6^2 C_4^2 C_2^2/3!=15$, соответствующих третьей форме {2}, {2}, {2}. Приведем теперь полный список разбиений, соответствующих всем видам распределений.

Форма распределения {4}, {1}, {1}

(1, 2, 3, 4), (5), (6)	(6, 2, 3, 4), (5), (1)
(1, 2, 3, 5), (4), (6)	(1, 2, 5, 6), (3), (4)
(1, 2, 5, 4), (3), (6)	(1, 5, 6, 4), (2), (3)
(1, 3, 5, 4), (2), (6)	(5, 6, 3, 4), (1), (2)
(5, 2, 3, 4), (1), (6)	(5, 2, 3, 6), (1), (4)
(1, 2, 3, 6), (5), (4)	(1, 5, 3, 6), (2), (4)
(1, 2, 6, 4), (5), (3)	(5, 2, 6, 4), (1), (3)
(1, 6, 3, 4), (5), (2)	

Форма распределения {3}, {2}, {1}

(1, 2, 3), (4, 5), (6)	(1, 4, 5), (2, 3), (6)
(1, 2, 3), (4, 6), (5)	(1, 4, 5), (2, 6), (3)
(1, 2, 3), (5, 6), (4)	(1, 4, 5), (3, 6), (2)
(1, 2, 4), (3, 5) (6)	(1, 4, 6), (2, 3), (5)
(1, 2, 4), (3, 6), (5)	(1, 4, 6), (2, 5), (3)
(1, 2, 4), (5, 6), (3)	(1, 4, 6), (3, 5), (2)
(1, 2, 5), (4, 3), (6)	(1, 5, 6), (2, 3), (4)
(1, 2, 5), (4, 6), (3)	(1, 5, 6), (2, 4), (3)
(1, 2, 5), (6, 3), (4)	(1, 5, 6), (3, 4) (2)
(1, 2, 6), (4, 5), (3)	(2, 4, 5), (1, 3), (6)
(1, 2, 6), (3, 5), (4)	(2, 4, 5), (1, 6), (3)
(1, 2, 6), (3, 4), (5)	(2, 4, 5), (3, 6), (1)
(1, 4, 3), (2, 5), (6)	(2, 4, 6), (1, 3), (5)
	(2, 4, 6), (1, 5), (3)
(1, 4, 3), (2, 6), (5)	(2, 4, 6), (3, 5), (1)
(1, 4, 3), (5, 6), (2)	(2, 5, 6), (1, 3), (4)
(1, 5, 3), (4, 2), (6)	(2, 5, 6), (1, 4), (3)
(1, 5, 3), (4, 6), (2)	(2, 5, 6), (3, 4), (1)
(1, 5, 3), (2, 6), (4)	(3, 4, 5), (1, 2), (6)
(1, 6, 3), (4, 5), (2)	(3, 4, 5), (1, 6), (2)
(1, 6, 3), (4, 2) (5)	(3, 4, 5), (2, 6), (1)
(1, 6, 3), (2, 5), (4)	(3, 4, 6), (1, 2), (5)
(4, 2, 3), (1, 5), (6)	(3, 4, 6), (1, 5), (2)
(4, 2, 3), (1, 6), (5)	(3, 4, 6), (2, 5), (1)
(4, 2, 3), (5, 6), (1)	(3, 5, 6), (1, 2), (4)
(5, 2, 3), (4, 1), (6)	(3, 5, 6), (1, 4), (2)
(5, 2, 3), (4, 6), (1)	(3, 5, 6), (2, 4), (1)
(5, 2, 3), (1, 6), (4)	

Продолжение

(6, 2, 3), (4, 5), (1)	(4, 5, 6), (1, 2), (3)
(6, 2, 3), (4, 1), (5)	(4, 5, 6), (1, 3), (2)
(6, 2, 3), (1, 5), (4)	(4, 5, 6), (2, 3), (1)

Форма распределения {2}, {2}, {2}

(1, 2), (3, 4), (5, 6)	(1, 4), (2, 6), (3, 5)
(1, 2), (3, 5), (4, 6)	(1, 5), (3, 4), (2, 6)
(1, 2), (3, 6), (4, 5)	(1, 5), (3, 2), (4, 6)
(1, 3), (2, 4), (5, 6)	(1, 5), (3, 6), (2, 4)
(1, 3), (2, 5), (4, 6)	(1, 6), (3, 4), (5, 2)
(1, 3), (2, 6), (4, 5)	(1, 6), (3, 5), (4, 2)
(1, 4), (2, 3), (5, 6)	(1, 6), (3, 2), (4, 5)
(1, 4), (2, 5), (3, 6)	

При полном переборе целевую функцию (ВСК) необходимо вычислить для каждой из 90 альтернатив разбиения, приведенных выше; затем отыскивается такое разбиение, которое приводит W к минимуму. Из приведенного списка возможных альтернатив разбиения видно, что ВСК для некоторых кластеров, например для (1, 2, 3), будет вычисляться больше одного раза.

Применение методов динамического программирования к проблеме кластеризации представляет собой последовательное нахождение оптимальной группировки, на каждом шаге которой вычисляется целевая функция; при этом лишние вычисления, присутствующие при методике полного перебора, исключаются. Другими словами, оптимальное решение находится поэтапно. Подход с помощью динамического программирования ускоряет обработку массива информации.

Описанный выше пример рассмотрим в качестве иллюстраций применения динамического программирования. Прежде всего все альтернативы кластеризации классифицируются соответственно формам распределения. Напомним, что компоненты форм распределения располагаются в убывающем порядке. На первом шаге для каждой компоненты первой формы распределения вычисляется и запоминается соответствующая целевая функция. На втором шаге для кластеров, соответствующих первым двум компонентам формы распределения вычисляются новые значения целевых функций; при этом привлекается вся информация, полученная на первом шаге; таким образом, внутригрупповая сумма квадратов не вычисляется заново для каждого кластера, а ее значение берется из предыдущего шага.

Для иллюстрации решения с помощью динамического программирования рассмотрим табл. 3.1. Второй столбец таблицы соответствует первой компоненте формы распределения, т. е. кластерам, образованным на первом шаге. Число кластеров первого шага равно: $C_6^4+C_6^3+C_6^2=50$. Функция W на первом шаге вычисляется для каждых из пяти кластеров. На втором шаге мы будем иметь 2 кластера, соответствующих первым двум компонентам формы распределения, размеры этих кластеров будут {4} и {1}, или {3} и {2} или {2} и {2}. Таким образом, общее число объектов на втором шаге равно 5 или 4. Число способов получения 5 объектов равно: $C_6^4\,C_2^1+C_6^3\,C_3^2=90$. Число способов получения 4 объектов равно: $C_6^2\,C_4^2+C_6^3\,C_3^1=150$, а общее число объектов на втором шаге равно 240. На втором шаге существует $\frac{1}{2}\,C_6^2\,C_4^2=45$ способов образования кластеров с компонентами формы распределения {2}, {2}; это означает, что мы имеем 45 повторений. На третьем шаге необходимо к компонентам {3} и {1} из второго шага добавить еще {2}, что приведет к форме {3} {1} {2} или эквивалентно {3} {2} {1}. Таким образом, общее число способов разбиения на втором шаге равно:

$$C_6^4\,C_2^1+C_6^3\,C_3^2+\frac{1}{2}\,C_6^2\,C_4^2=30+60+45=135$$

вместо 105.

Число различных множеств, содержащих 4 или 5 объектов, на втором шаге равно: $C_6^5+C_6^4=21$. Эти множества приведены в табл. 3.1 (шаг 2) и называются *состояниями*. Итак, на втором шаге имеется 21 состояние. На первом шаге число состояний равно 50. В табл. 3.1 показано 5 из 135 *допустимых способов* получения состояний на втором шаге.

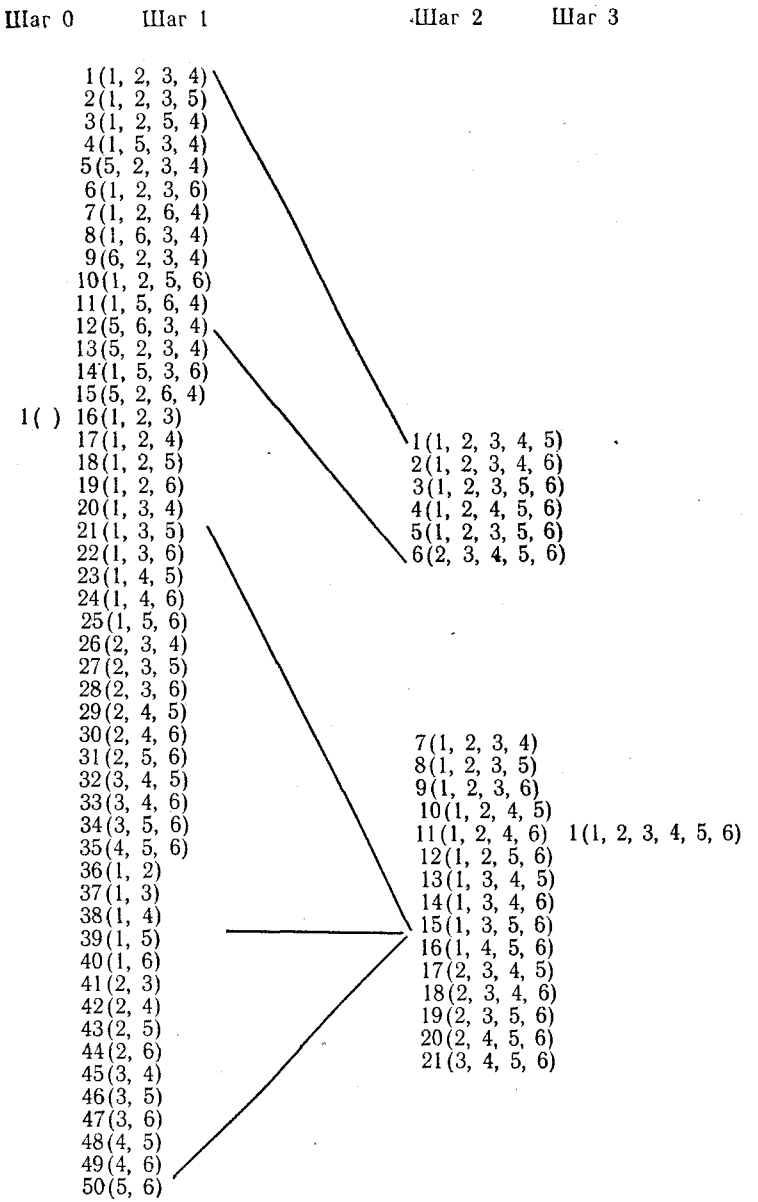
Третий шаг является окончательным. Он состоит из 3 кластеров. На последнем этапе имеется одно состояние, содержащее все 6 объектов. Число способов получения шести объектов на последнем шаге равно:

$$C_6^5\,C_1^1+C_6^4\,C_2^2=6+15=21,$$

т. е. имеется 21 допустимая дуга, связывающая шаг 2 с шагом 3.

Например, если $n=6$ и $m=3$, то общее число допустимых дуг равно: $135+21=156$. При включении первоначальных состояний число их будет: $156+50=206$.

Таблица 3.1



Каждая допустимая дуга приводит к *переходным вычислениям* по формуле

$$T(g_k)=\frac{1}{n_k}\sum_{i<j\in g_k} d_{ij}^2, \qquad (3.2)$$

где g_k обозначает группу из n_k объектов, а d_{ij} — расстояние между X_i и X_j.

В нашем примере всего 90 альтернатив разбиения, для каждой из которых необходимо произвести 3 переходные вычисления (всего 270). Решение с помощью динамического программирования требует 206 или по крайней мере 64 переходных вычислений.

Предположим, что на некотором шаге k имеется некоторое распределение объектов $X_1, \ldots, X_q$, $q \leqslant n$. В процедуре динамического программирования запоминается оптимальный способ разбиения q объектов на k непустых непересекающихся подмножеств. На следующих шагах, на которых требуется разбить q объектов на k групп, уже не будет необходимости решать эту задачу, перебирая все возможные способы их разбиения.

В качестве иллюстрации рассмотрим наш пример, в котором $n=6$ и $m=3$.

Таблица 3.2

Альтернатива	Переходные вычисления
1	$T(1,2)+T(3,4)+T(5,6)$
2	$T(1,3)+T(2,4)+T(5,6)$
3	$T(1,4)+T(2,3)+T(5,6)$

Напомним, что при $n=6$ и $m=3$ имеется $S(6,3)=90$ альтернатив разбиения. Три из 90 возможных альтернатив приведены в табл. 3.2. При полном переборе для трех альтернатив потребовалось бы 9 переходных вычислений. При оптимальном разбиении множества (1, 2, 3, 4) на две группы размера 2 требуется только 6 переходных вычислений. Если запомнить оптимальное разбиение $T(1,3)+T(2,4)=W_2(1,2,3,4)$, то для определения $W_2(1,2,3,4)+T(5,6)$ требуется только одно вычисление $T(5,6)$. Для трех альтернатив применение динамического программирования дает возможность, таким образом, сократить число вычислений на 9—7=2. Естественно, при увеличении n и m число сокращаемых вычислений значительно увеличивается, однако по отношению к общему числу переходных вычислений это увеличение может быть не столь значительным.

3.2. Модель динамического программирования Дженсена

В отличие от линейного программирования, для которого существует точная стандартная формулировка задачи, в динамическом программировании таковая отсутствует.

В задачах динамического программирования соответствующие формулы и уравнения зависят от конкретной постановки проблемы. При этом, как правило, задачу сводят к последовательности рекурсивных формул или уравнений, отражающих связи, имевших место в задаче, по которым и отыскивается «оптимальное» решение. Рао [291] приводит формулировку применения динамического программирования для решения кластерной проблемы при $p=1$. Дженсен [183] описывает более общую ситуацию, которую мы сейчас и рассмотрим.

На языке динамического программирования задачу кластеризации Дженсен описывает в виде следующих рекурсивных уравнений:

$$W_k(z)=\begin{cases} 0, \text{ если } k=0; \\ \min\limits_{y}[T(z-y)+W_{k-1}(y)], \\ \text{если } k=1,2,\ldots,m_0. \end{cases} \qquad (3.3)$$

где m — число непересекающихся непустых подмножеств, на которые разбивается n объектов;

k — индекс или переменная шага; $m_0=m$, если $n \geqslant m$, и равно $n-m$, если $n<m$;

z — переменная состояния, характеризующая данное множество объектов на шаге k;

y — переменная состояния, характеризующая данное множество объектов на шаге $k-1$;

$z-y$ — подмножество всех объектов, содержащихся в z и не содержащихся в y;

$T(z-y)$ — «переходные издержки» (transition cost) объектов в кластере $(z-y)$.

Переменные y и z представляют собой два состояния (множества объектов) на шаге $k-1$ и k соответственно.

Разность $z—y$ представляет собой те объекты, которые содержатся на шаге k и не содержатся на шаге $k—1$. $T(z—y)$ означает «переходные издержки», т. е. ВСК объектов, объединенных на шаге $k—1$; $W_k(z)=\min_y [T(z—y)+W_{k-1}(y)]$ дает минимальное значение ВСК при разбиении объектов, содержащихся в z, на k непустых непересекающихся подмножеств. Очевидно, формула (3.3) приводит к большому числу вычислений. Напомним читателю, что, как следует из уравнения (3.2) параграфа 3.1, где g_i обозначает кластер из n_i объектов, переходные издержки $T(g_i)$ равны:

$$T(g_i)=\frac{1}{n_i}\sum_{k<j\in g_i} d_{kj}^2,$$

что совпадает в ВСК кластера g_i.

Заметим, что число шагов равно $m_0=m$, если $n\geqslant m$, и $n—m$, если $n<2m$. Это объясняется тем, что если $n<2m$, то обязательно существуют по крайней мере $n—m+1$ кластеров, состоящих из одного объекта. Переходные издержки T для кластера из одного объекта равны нулю, поэтому вклад такого кластера в W также равен нулю. Следовательно, процесс может закончиться на шаге m_0; при этом все оставшиеся кластеры будут содержать по одному объекту. При вычислении $W_k(z)$ необходимо подчеркнуть, что объекты, соответствующие любому состоянию на шаге k, состоят из объектов некоторого множества, соответствующего некоторому состоянию y на шаге $k—1$, и объектов, содержащихся в другом множестве, содержащемся в $z—y$.

В качестве примера, иллюстрирующего рекурсивные уравнения (3.3), рассмотрим 37-е состояние первого шага и 15-е состояние второго шага; напомним, что $n=6$, $m=3$ (табл. 3.1). В этом случае y обозначает объекты (1, 3) 37-го состояния первого шага, z — объекты (1, 3, 5, 6) 15-го состояния второго шага, а $z—y$ — объекты (5, 6). «Переходными издержками» из 37-го состояния в 15-е будут:

$$T(z-y)=T(5,6)=d_{56}^2.$$

Переходными издержками из 37-го состояния первого шага в первое состояние второго шага будут:

$$T(z-y)=T(2,4,5)=\frac{d_{24}^2+d_{25}^2+d_{45}^2}{3}.$$

На первом шаге алгоритма динамического программирования для данного множества кластеров вычисляется значение $W_1(z)$. В этом случае

$$W_1(z)=\min_y [T(z-y)+W_0(y)]=T(z),$$

где z обозначает данное множество объектов. Значение $W_1(z)$ вычисляется для каждого из кластеров первого шага. Максимальное число объектов, содержащихся в кластере на первом шаге, обозначим через max (1):

$$\max(1)=n-m+1;$$

это означает, что максимальный кластер содержит $n—m+1$ объектов, а все остальные кластеры по одному. Минимальное число объектов в кластере на первом шаге, которое обозначим через min (1), равно:

$$\min(1)=n/m,$$

если n делится нацело на m, и

$$\min(1)=\begin{cases}[n/m]+1 & \text{для } 1\leqslant n-m[n/m],\\ n-(m-1)[n/m] & \text{для } n-m[n/m]<1\leqslant m,\end{cases}$$

если n не делится нацело на m; $[n/m]$ обозначает наибольшее число, меньшее или равное n/m. Общее число кластеров на первом шаге, которое обозначим через $NS(1)$, равно:

$$NS(1)=\sum_{j=\min(1)}^{\max(1)} C_n^j. \qquad (3.4)$$

На первом шаге алгоритма для всех возможных кластеров $NS(1)$ вычисляется значение $T(z)$.

В общем случае максимальное число объектов в одном состоянии на шаге k равно максимальной сумме компонент всех форм распределения с первого по k-й шаг включительно. Минимальное число состояний равно минимальной сумме компонент форм распределения. $\max(k)$ и $\min(k)$ определяются из следующих уравнений:

$$\max(k)=n-m+k \qquad (3.5)$$

и

$$\min(k)=k[n/m], \qquad (3.6)$$

если n делится на m нацело. В противном случае

$$\min(k)=\begin{cases}([n/m]+1)k \text{ для } 1\leqslant k\leqslant n-m[n/m],\\ n-(m-k)[n/m] \text{ для } n-m[n/m]<\\ \qquad <k\leqslant m.\end{cases} \quad (3.7)$$

Число состояний на шаге k определяется как

$$NS(k)=\begin{cases}1 \text{ для } k=0,\\ \sum\limits_{j=\min(k)}^{\max(k)} C_n^j \text{ для } k=1,2,\ldots,m_0.\end{cases} \quad (3.8)$$

Общее число состояний при динамическом программировании равно

$$\sum_{k=0}^{m_0} NS(k). \quad (3.9)$$

В рассмотренном подходе большое значение имеет общее число значений $W_k(z)$ при переходе от шага k—1 к шагу k, т. е. число способов построения состояния на шаге k. Состояния последовательных шагов связываются дугами. Два состояния на шаге k и k—1 связываются допустимыми дугами, если объекты состояния на шаге k связаны с состоянием на шаге k—1. Отсюда следует, что допустимая дуга не может связывать состояние на шаге k—1 и состояние на шаге k, если объект некоторого состояния на шаге k—1 не связан ни с одним состоянием на шаге k для $2\leqslant k\leqslant m_0$.

В алгоритме динамического программирования общее число допустимых дуг равно:

$$TFA=NS(1)+\sum_{k=1}^{m_0-1} TA(k), \quad (3.10)$$

где $TA(k)$ обозначает общее число дуг между шагами k и $k+1$ для $k=1,2,\ldots,m_0$. Значение $TA(k)$ находится по формуле

$$TA(k)=\sum_{j=\min(k)}^{\max(k)}\ \sum_{i=1}^{\max(k+1)-\min(k)} FA(j,i), \quad (3.11)$$

где

$$FA(j,i)=\begin{cases}C_n^j C_{n-j}^i, \text{ если } \min(k+1)\leqslant i+j\leqslant\\ \qquad\qquad \leqslant\max(k+1),\\ 0 \text{ в других случаях.}\end{cases} \quad (3.12)$$

В уравнениях (3.11) и (3.12) i обозначает число объектов в классе *(допустимых) состояний* шага k. Всего имеется C_n^i таких состояний, содержащих i объектов; это следует из того факта, что число подмножеств, содержащих i объектов, равно C_n^i. Символом j обозначено число объектов, которые комбинируются с i объектами и образуют новое состояние шага $k+1$. Очевидно, для состояния порядка $i+j$, присутствующего на шаге $k+1$, $\min(k+1)\leqslant i+j\leqslant\max(k+1)$. Если $i+j$ удовлетворяет приведенному условию, то существует C^j_{n-i} множеств порядка n—i, которые добавляются к i объектам k раз.

Дженсен предлагает способ вычисления эффективности динамического программирования по сравнению с методом полного перебора. Эта эффективность измеряется отношением общего числа переходных вычислений при динамическом программировании к соответствующему числу вычислений при полном переборе. В качестве альтернативы в числителе можно взять общее число допустимых дуг. В любом случае процедура динамического программирования довольно эффективна. Однако применение методов динамического программирования требует привлечения дополнительной памяти ЭВМ; соответственно увеличивается время вычислений за счет обращения к памяти машины, что в конечном счете может привести к тому, что метод полного перебора окажется более предпочтительным*. В любом случае для больших n и m, может быть, целесообразнее воспользоваться другими методами, например ISODATA [18] или ступенчатым.

Для иллюстрации методики, предложенной Дженсеном, рассмотрим наш пример, в котором $n=6$, а $m=3$. В этом случае $n=2m$ и поэтому нам необходимо рассмотреть n—m=6—3=3 шага, т. е. $m_0=3$. Кроме того,

$$\begin{aligned}&\max(1)=n-m+1=4; \quad &&\min(2)=([6/3])2=4;\\ &\min(1)=([6/3])1=2; \quad &&\max(3)=n-m+3=6;\\ &\max(2)=n-m+2=5; \quad &&\min(3)=([6/3])3=6,\end{aligned}$$

как следует из табл. 3.1.

---

* Как известно, время обращения к дополнительной памяти ЭВМ намного больше времени вычислений с данными из оперативной памяти. — *Примеч. пер.*

Общее число состояний на шаге 0, 1, 2 и 3 находятся из формул:

$$NS(0)=1;$$

$$NS(1)=C_6^4+C_6^3+C_6^2=50;$$

$$NS(2)=C_6^4+C_6^5=21;$$

$$NS(3)=C_6^6=1.$$

Эти состояния приведены в табл. 3.1. Таким образом, общее число состояний равно 73. Это число можно было бы получить непосредственным подсчетом состояний в табл. 3.1 ($NS(0)=1$).

Для $k=1$ из уравнения (3.12) находим:

$$FA(3,1)=C_6^3C_3^1=60; \qquad FA(2,2)=C_6^2C_4^2=90;$$

$$FA(3,2)=C_6^3C_3^2=60; \qquad FA(3,3)=FA(4,2)=0$$

$$FA(4,1)=C_6^4C_2^1=30;$$

и общее число допустимых дуг между шагами 1 и 2 равно $TA(1)=240$.

Для $k=2$ получим:

$$TA(2)=C_6^4C_2^2+C_6^5C_1^1=21.$$

Таким образом, общее число допустимых дуг в нашем примере, как следует из (3.10), равно:

$$TFA=NS(1)+\sum_{k=1}^{2}TA(k)=50+240+21=311.$$

В параграфе 3.1 было показано, что половина состояний на шаге {2}, соответствующих компонентам форм распределения {2} {2}, лишние, и 60 дуг, соответствующие компонентам {3} {1}, в конечном счете приводят к форме распределения {3} {1} {2}, которая эквивалентна форме {3} {2} {1}. Таким образом, при редуцировании число допустимых дуг, которое обозначим через NA, равно: $NA=50+135+21=206$.

Число допустимых дуг между шагами k и $k+1$ после исключения равно:

$$NA(k)=\sum_{i=\min(k)}^{\max(k)}\ \sum_{j=1}^{\max(k+1)-\min(k)}A(i,j),$$

где

$$A(i,j)=\left\{\begin{array}{l} C_n^i\,C_{n-i}^j, \text{ если } i\neq j,\\ \frac{1}{2}C_n^i\,C_{n-i}^j, \text{ если } i=j,\\ 0 \text{ в других случаях.}\end{array}\right\}$$

и $$\min(k+1)\leqslant i+j\leqslant\max(k+1),$$
$$(m-k)j+i\geqslant n.$$

Общее число дуг после редуцирования равно:

$$NA=NS(1)+\sum_{k=1}^{m_0-1}NA(k). \qquad (3.13)$$

Легко проверить, что при $n=6$ и $m=3$ уравнение (3.13) приведет к значению $NA=206$. Итак, максимальное число допустимых дуг, которые необходимо рассмотреть при динамическом программировании, для нашего примера равно 206.

Чтобы показать, как работает алгоритм динамического программирования, допустим $p=2$, а шесть объектов равны (1, 1), (3, 4), (5, 5), (4, 4), (1, 2) и (5, 6), т. е.

$$X=\begin{pmatrix}1&3&5&4&1&5\\1&4&5&4&2&6\end{pmatrix}.$$

Квадраты расстояний равны:

$$d_{12}^2=13,\ d_{13}^2=32,\ d_{14}^2=18,\ d_{15}^2=1,\ d_{16}^2=41,$$

$$d_{23}^2=5,\ d_{24}^2=1,\ d_{25}^2=8,\ d_{26}^2=8,\ d_{34}^2=2,$$

$$d_{35}^2=25,\ d_{36}^2=1,\ d_{45}^2=13,\ d_{46}^2=5,\ d_{56}^2=32.$$

Следуя алгоритму динамического программирования, будем иметь:

шаг 0. $W_0(0)=0$;

шаг 1. Вычислим $W_1(z)=T(z-y)+W_0(y)=T(z)+W_0(0)=T(z)$ для каждого множества объектов шага 1. Например,

$$W_1(1,2,3,4)=T(1,2,3,4)+W_0(0)=$$
$$=\frac{(d_{12}^2+d_{13}^2+d_{14}^2+d_{23}^2+d_{24}^2+d_{34}^2)}{4}=17{,}75.$$

На данном шаге мы должны получить 50 таких значений;

шаг 2. Для каждого множества объектов этого шага вычислим $W_2(z) = \min\limits_y \{T(z-y) + W_1(y)\}$.

Например,

$$W_2 = (1, 2, 3, 4, 5) = \min\{T(5) + W_1(1, 2, 3, 4),$$
$$T(4) + W_1(1, 2, 3, 5),$$
$$T(3) + W_1(1, 2, 4, 5), T(2) + W_1(1, 3, 4, 5),$$
$$T(1) + W_1(2, 3, 4, 5), T(1, 2) + W_1(3, 4, 5),$$
$$T(1, 3) + W_1(2, 4, 5), T(1, 4) + W_1(2, 3, 5),$$
$$T(1, 5) + W_1(2, 3, 4), T(2, 3) + W_1(1, 4, 5),$$
$$T(2, 4) + W_1(1, 3, 5), T(2, 5) + W_1(1, 3, 4),$$
$$T(3, 4) + W_1(1, 2, 5), T(3, 5) + W_1(1, 2, 4),$$
$$T(4, 5) + W_1(1, 2, 3)\};$$

шаг 3. Для каждого множества объектов этого шага вычислим $W_3(z) = \min\limits_y \{T(z-y) + W_2(y)\}$. На этом шаге z обозначает единственное множество объектов (1, 2, 3, 4, 5, 6). Всего существует 21 допустимая дуга, соединяющая состояния шага 2 с состоянием шага 3. Таким образом, нам необходимо выбрать, как минимум, 21 значение. Одно из этих значений равно

$$T(2, 4) + W_2(1, 3, 5, 6).$$

Оно соответствует состоянию под номером 15 (см. табл. 3.1), для которого y соответствует множеству (1, 3, 5, 6), z соответствует (1, 2, 3, 4, 5, 6), а $z-y$ множеству (2, 4).

Результатом применения процедуры динамического программирования будут кластеры (1, 1) и (1, 2), (3, 4) и (4, 4), (5, 5) и (5, 6) с формой распределения {2}, {2}, {2}. Максимальное значение равно:

$$W_3(1, 2, 3, 4, 5, 6) = 1{,}5.$$

Решение показано на рис. 2.

3.3. Применение целочисленного программирования в кластерном анализе

В этом параграфе кластерная проблема будет рассмотрена в терминах целочисленного программирования; впервые это было сделано Вайнодом [380] и Рао [291].

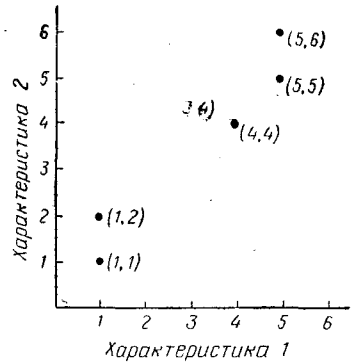


Рис. 2. Граф для $n=6$ объектов

Балинский [12] в своем обзоре рассматривает новые методы целочисленного программирования. Посоветуем читателю также работы [11], [26], [134] и [228].

В формулировке Вайнода [380] рассматривается случай одной измеряемой характеристики ($p=1$), в соответствии с которой и выполняется разбиение на кластеры*. Предположим, у нас имеется n объектов со значениями $x_1, \ldots, x_n$. Обозначим через n_j число объектов в j-й группе ($j=1,2, \ldots, m$); $n = \sum\limits_{j=1}^{m} n_j$. Число объектов в максимальном (по числу их) кластере обозначим через m_0. Если m_0 специально не определено, то $m_0=n$. «Издержки», связанные с включением i-го объекта в j-ю группу, обозначим через c_{ij}, очевидно, $c_{ii}=0$, $c_{ij}=c_{ji}$. Положим $a_{ij}=1$, если i-й объект принадлежит j-й группе, и $a_{ij}=0$ — в противном случае. Имеем $n_j = \sum\limits_{i=1}^{n} a_{ij}$, т. е. n_j есть сумма элементов в j-м столбце матрицы $A=\{a_{ij}\}$.

Если число кластеров m известно заранее, то матрицы $C=\{c_{ij}\}$ и $A=\{a_{ij}\}$ будут иметь порядок $m \times n$. Однако, если m неизвестно, порядок матриц C и A также неизвестен. Чтобы обойти это ограничение, предположим, что общее число групп равно n, а число пустых групп равно $n-m$. В целях идентификации групп введем понятие так называемого *ведущего элемента группы*. По-

* В этом случае задача сводится к простой группировке по одному признаку. — *Примеч. ред.*

ложим $y_j=1$, если j-й объект является ведущим, и $y_j=0$ в противном случае.

В формулировке кластерной задачи на языке целочисленного программирования матрица издержек C предполагается известной. Например, в качестве c_{ij} может быть рассмотрено приращение в ВСК при включении i-го объекта в группу с j-м ведущим элементом. Задача целочисленного программирования заключается в минимизации общей суммы издержек по всем группам разбиения при условии некоторых ограничений:

$$\text{минимизировать} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}c_{ij}$$

при условии, что

1) $\sum_{i=1}^{n} a_{ij}=1,\ i=1,2,\ \ldots,\ n;$ (3.14)

2) $\sum_{j=1}^{n} y_j=m;$ (3.15)

3) $y_j \geqslant a_{1j},\ y_j \geqslant a_{2j},\ \ldots,\ y_j \geqslant a_{nj},\ j=1,2,\ \ldots,\ n.$ (3.16)

Ограничение (1) означает, что никакой объект не может содержаться более чем в одной группе. Ограничение (2) означает, что существует ровно m групп. Последнее условие означает, что, прежде чем объекты будут присоединены к группе, содержащей j-й элемент, этот элемент должен быть ведущим. Ограничение (3) можно переписать следующим образом:

3)′ $ny_j \geqslant \sum_{i=1}^{n} a_{ij},\ j=1,2,\ \ldots,\ n.$

Приведенная формулировка может быть применена и для случая, когда каждый объект имеет p измерений, т. е.

$$X_i^T = (x_{1i},\ x_{2i},\ \ldots,\ x_{pi}).$$

В этом случае в качестве издержек может выступать любая метрика расстояния из главы 1. Порядок матрицы C при этом не меняется; задача заключается в минимизации C с учетом его нового определения.

Во *второй формулировке* Вайнода издержки c_{ij} определяются в терминах ВСК. Значения $x_1,\ x_2,\ \ldots,\ x_n$ для n объектов располагаются в возрастающем порядке $z_1, z_2, \ldots, z_n$, т. е.

$$z_1 \leqslant z_2 \leqslant \ldots \leqslant z_n.$$

Проблема заключается в разбиении $z_1,\ z_2,\ \ldots,\ z_n$. Обо-

значим через g_j группу, минимальное значение которой равно z_j. Элемент z_j назовем ведущим элементом группы g_j. Матрица A определяется так же, как и в предыдущем случае, т. е. $a_{ij}=1$ или 0 в зависимости от того, принадлежит ли z_i к g_j или нет. Поскольку z упорядочены, z_i не может принадлежать к $g_{i+1},\ \ldots,\ g_n$, это означает, что элементы матрицы A, которые лежат выше диагонали, равны нулю, или

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} a_{ij}=0.$$

Задачу теперь можно представить как минимизацию

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}(z_i-\overline{z_j})^2, \qquad (3.17)$$

где $\overline{z_j}$ есть среднее j-й группы

$$\overline{z_j}=\sum_{i=1}^{n} a_{ij}z_i/n_j$$

и

$$\sum_{i=1}^{n} a_{ij}=n_j\ .$$

Для того чтобы (3.17) обращалось в минимум, необходимо выполнение так называемого условия цепи (string property). Условие цепи означает, что не существует группы, содержащей z_i и z_j $(i<j)$ и не содержащей все z, значения которых лежат между ними. Это означает, что матрица $A=\{a_{ij}\}$ не может содержать прерывающихся цепочек из единиц, расположенных ниже главной диагонали, т. е.

$$a_{jj} \geqslant a_{j+1,j} \geqslant \ldots \geqslant a_{nj},\ j=1,2,\ \ldots,\ n.$$

Максимальная цепочка содержит m_0 членов, поскольку максимальное число элементов, которое может содержать группа, равно m_0.

Лемма 3.1. Условие цепи $a_{jj} \geqslant a_{j+1,\,j} \geqslant \ldots \geqslant a_{nj}$ является необходимым условием минимальности.

Доказательство. При объединении объектов z_j, $z_{j+1},\ \ldots,\ z_{j+k-1}$ с z_{j+k} увеличение ВСК равно следу матрицы (1.10), т. е.

$$\frac{k}{k+1}\left(z_{j+k}-\frac{1}{k}\sum_{i=0}^{k-1} z_{j+i}\right)^2 .$$

То же происходит при объединении z_j, z_{j+1}, ..., z_{j+k-1} с z_{j+k+1}; увеличение равно

$$\frac{k}{k+1}\left(z_{j+k+1}-\frac{1}{k}\sum_{i=0}^{k-1} z_{j+i}\right)^2 .$$

Однако $z_{j+k+1}>z_{j+k}$, что и доказывает необходимый результат.

Минимизация $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(z_i-\bar{z}_j)^2$ эквивалентна минимизации $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}c_{ij}$ при соответствующем выборе $A=\{a_{ij}\}$ и $C=\{c_{ij}\}$:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}c_{ij}=\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(z_i-\bar{z}_j)^2 .$$

Значения c_{ij} необходимо определить только для $i>j$, т. е. для элементов ниже главной диагонали. Значение c_{ij} численно равно приращению ВСК при включении элемента z_i в группу, состоящую из элементов z_j, ..., z_{i-1}. Принимая во внимание (1.10), имеем:

$$c_{ij}=\frac{i-j}{i-j+1}\left(z_i-\frac{1}{i-j}\sum_{k=j}^{i-1} z_k\right)^2 . \qquad (3.18)$$

При таком определении c_{ij} можно показать, что

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}c_{ij}=\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(z_i-\bar{z}_j)^2,$$

с помощью которого первоначальную задачу квадратичного программирования можно свести к задаче линейного программирования. Окончательно задачу можно сформулировать следующим образом:

$$\text{минимизировать } \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}c_{ij}$$

при условии:

1) $a_{jj}>a_{j+1,\,j}>\ldots>a_{nj}$, $j=1, 2, \ldots, n$;

2) $c_{ij}=\frac{i-j}{i-j+1}\left(z_i-\frac{1}{i-j}\sum_{k=j}^{i-1} z_k\right)^2$.

Обобщение второй формулировки Вайнода на многомерный случай более затруднительно, чем первой. Пусть имеется X_1, X_2, ..., X_n точек в p-мерном пространстве. Поскольку условие цепи нельзя непосредственно обобщить с одномерного случая на многомерный, Вайнод [380] предлагает обобщенное условие цепи, которое можно применить и в многомерном случае. Им также установлено, что условие цепи — это необходимое условие минимума ВСК. Рао [291] приводит контрпример и предлагает альтернативную форму обобщенного условия цепи, которое также представляет собой необходимое условие минимума ВСК.

Мы приведем оба определения после чего коротко обсудим два соответствующие им варианта проблемы (Рао и Вайнода). Пусть $D=\{d_{ij}\}$ обозначает матрицу $n\times n$ парных евклидовых расстояний. Элементы j-го столбца перегруппируем в порядке возрастания, т. е.

$$d_{jj}\leqslant d_{i_1,\,j}\leqslant\ldots\leqslant d_{i_{n-1},\,j},$$

что совпадает с одномерным случаем для z. В терминах матрицы A обобщенное условие цепи означает, что цепь из единиц должна проходить через $i_1, i_2, \ldots, i_{n-1}$. Максимальная длина цепи для любого столбца равна m_0.

Приведем теперь два определения обобщенного условия цепи.

1. (Вайнод). Цепь из единиц для j-го столбца матрицы A проходит через $i_1, i_2, \ldots, i_{n-1}$, причем

$$d_{jj}=0\leqslant d_{i_1,j}\leqslant d_{i_2,j}\leqslant\ldots\leqslant d_{i_{n-1},j} .$$

2. (Рао). Каждая группа содержит хотя бы один объект (ведущий группы), такой, что расстояние между этим объектом и любым другим объектом, не входящим в эту группу, не меньше, чем расстояние между этим объектом и любым объектом из этой группы.

$c_{i_k,\,j}$ $(i_k\neq j)$ Вайнод определяет как приращение ВСК при включении X_{i_k} в группу из объектов j, i_1, ..., i_{k-1}. Значение $c_{i_k,\,j}$ аналогично (3.18) и в многомерном случае равно:

$$c_{i_k,j}=\frac{k}{k+1}\left(d^2_{i_k,j}-\frac{1}{k}\sum_{m=1}^{k} d_{i_m,j}\right)^2 . \qquad (3.19)$$

Пользуясь матрицей издержек, соответствующей уравнению (3.19), и матрицей A, элементы которой удовле-

творяют неравенству $a_{ij} \geqslant a_{i_1, j} \geqslant a_{i_2, j} \geqslant \ldots \geqslant a_{i_{n-1}, j}$, задачу минимизации ВСК можно свести к минимизации $\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} c_{ij}$.

При формулировке задачи линейного целочисленного программирования Рао [291] гребует выполнения условия цепи 2. Задача, таким образом, заключается в том, чтобы найти

$$\min C^T Y \qquad (3.20)$$

при условии $AY = b^T$,

где $y_i = 0$ или 1 (за исключением последнего элемента);

A — матрица $(n+1) \times [n(n-1)+2]$;

Y — вектор-столбец порядка $n(n-1)+2$;

b — вектор-столбец порядка $n+1$, равный $(1,1, \ldots, 1, m)^T$;

C^T — вектор-строка порядка $n(n-1)+2$.

Заметим, что минимизируемая величина (3.20) снова является *линейной* функцией переменных, состоящих из 0 и 1. Каждый элемент Y служит индикатором группы, т. е. $y_i = 0$ или 1 в зависимости от того, пользуются ли i-й группой в окончательном решении или нет. Вектор C^T является вектором значений целевой функции; c_i представляет собой значение целевой функции i-й группы. Например, если в качестве целевой функции рассматривается ВСК, то c_i имеет вид:

$$c_i = \sum_{l=1}^{ni} \sum_{j=1}^{p} (x_{jl} - \bar{x}_j)^2 .$$

Вектор Y содержит m единиц и $n(n-1)+2-m$ нулей. Выражение $C^T Y$ определяет значение целевой функции для данной альтернативы разбиения.

Матрица A аналогична соответствующей матрице из (3.17); однако теперь ее порядок равен $(n+1)[n(n-1)+2]$. Матрица A удовлетворяет следующим условиям:

1) последняя строка A состоит из единиц, т. е.

$$a_{n+1, j} = 1, \; j = 1, 2, \ldots, n(n-1)+2;$$

2) каждая строка A, за исключением последней, соответствует одному объекту и содержит только одну единицу; все остальные $n(n-1)+1$ элементы равны нулю;

3) каждый столбец A, за исключением последнего, представляет собой коэффициенты группы, т. е.

$$\sum_{i=1}^{n} a_{ij} = n_j, \quad j = 1, 2, \ldots, n(n-1)+1;$$

4) последний столбец A состоит из нулей, за исключением последнего элемента, который равен 1.

Последний элемент в векторе b равен общему числу кластеров, т. е. m.

Будем предполагать, что $d_{ij} \neq d_{ik}$, если $j \neq k$. Из условия цепи следует, что существует $n-1$ групп, в которых j-й объект является ведущим. Для демонстрации этого факта рассмотрим неравенство

$$d_{jj} = 0 < d_{i_1, j} < d_{i_2, j} < \ldots < d_{i_{n-1}, j} \cdot .$$

При условии, что j-й объект является ведущим, возможны следующие разбиения на группы: (j, i_1), (j, i_1, i_2), ..., ..., $(j, i_1, i_2, \ldots, i_{n-1})$. Число таких групп равно $n-1$. Поскольку всего имеется n объектов, то отсюда следует, что общее число возможных групп равно $n(n-1)+1$, включая и ту группу, которая содержит все объекты. Число элементов в векторе Y соответствует общему числу возможных кластеров с тем ограничением, что общее число возможных кластеров равно m. Задача (3.20) вместе с соответствующими ограничениями может быть решена различными методами, описанными в работе [127].

Рао приводит другие критерии минимизации, которыми также можно пользоваться и которые приводят к задаче математического программирования. Это

1) минимизация суммы средних квадратов внутригрупповых расстояний;

2) минимизация общей суммы внутригрупповых расстояний;

3) минимизация максимума внутригрупповых расстояний.

С вычислительной точки зрения эти критерии неудобны, за исключением случаев небольших значений n и m. Если число групп $m=2$, то они более удобны. Это значение m довольно популярно, если иметь в виду метод Эдвардса и Кавалли-Сфорца [93], который делит все объекты на две компактные группы, и процедура повторяется.

ГЛАВА 4

ПРЕДСТАВЛЕНИЯ МАТРИЦ СХОДСТВ

Как было отмечено в предыдущих главах, задачи кластерного анализа можно решать как в терминах матрицы расстояний D, так и в терминах матрицы сходства S. В этой главе мы рассмотрим различные аспекты представления результатов кластеризации, матриц сходства и расстояния.

Параграфы 4.1 и 4.2 служат неформальным введением в более строгое изложение соответствующих вопросов с точными формулировками (Хартиген), которые читатель найдет в параграфе 4.3.

4.1. Дендограммы

Последовательный процесс кластеризации начинается с рассмотрения n объектов; затем два наименее удаленных (ближайших) объекта объединяются в один кластер и число кластеров становится равным $n-1$. Процесс повторяется до тех пор, пока все n объектов не попадут в один кластер, содержащий все объекты. Идеи, изложенные в этой главе, имеют в виду применение методов последовательной (стратифицированной) кластеризации.

Наиболее известный метод представления матрицы расстояний (разнородности) или сходства основан на идее «дендограммы», или «диаграммы-дерева».

Дендограмму можно определить как графическое изображение результатов процесса последовательной кластеризации, который осуществляется в терминах матрицы расстояний или сходства. В дальнейшем процесс такой кластеризации будем рассматривать, как процедуру с матрицей расстояний или сходства. Таким образом, с помощью дендограммы можно графически

или геометрически изображать процедуру кластеризации при условии, что эта процедура оперирует только с элементами матрицы расстояний или сходства.

Существует много способов построения диаграмм-деревьев, соответствующих данной дендограмме. В диаграмме-дереве объекты располагаются вертикально слева, а результаты кластеризации справа. Значения расстояний или сходства, отвечающие построению новых кластеров, изображаются на горизонтальной прямой поверх дендограммы. Имея n объектов, можно построить большое количество диаграмм-деревьев, которые соответствуют данной процедуре кластеризации, однако для данной конкретной матрицы расстояний или сходства существует только одна диаграмма-дерево.

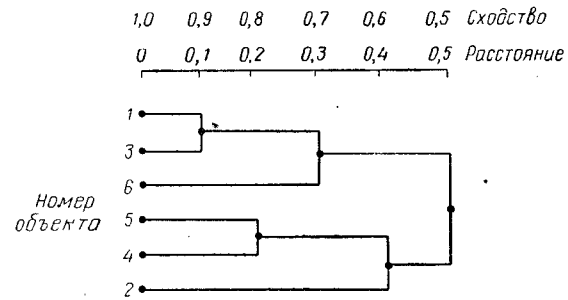


Рис. 3

На рис. 3 показан один из примеров диаграммы-дерева. В дальнейшем мы будем рассматривать диаграммы-деревья только одного вида, и поэтому дендограммы и диаграммы-деревья будут отождествляться. Рис. 3 соответствует случаю шести объектов ($n=6$) и p характеристик (признаков). Объекты 1 и 3 наиболее близки (наименее удалены друг от друга), и поэтому объединяются в один кластер на уровне близости, равном 0,9. Объекты 4 и 5 объединяются при уровне 0,8. На этом шаге имеются 4 кластера: (1, 3), (6), (5, 4) (2). На третьем и четвертом шаге процесса образуются кластеры (1, 3, 6) и (5, 4, 2), соответствующие уровню близости, равному 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне 0,5. С некоторыми специальными вопросами представления результатов кластеризации читатель может ознакомиться по работе

Сокала и Снита [336]. Вид дендограммы зависит от выбора меры сходства или расстояния между объектом и кластером и методом кластеризации. Наиболее важным моментом является выбор меры сходства или меры расстояния между объектом и кластером. Некоторые меры расстояния обсуждались в главе 1. Дендограмму можно построить для любой из этих мер.

Джонсон [186] рассматривает различные последовательные процедуры кластеризации и их связь с одной специальной метрикой. Рассмотрим последовательность множества кластеров C_0, C_1, $C_2, \ldots$, C_m и связанные с ними числа $\alpha_j = 0$, $j = 0, 1, \ldots m$. В терминах примера, показанного на рис. 3, C_j соответствует *точке ответвления*, α_j — уровню, при котором производится кластеризация. Множество C_0 содержит n кластеров, состоящих из одного объекта, а $\alpha_0 = 0$. Таким образом, $\alpha_0 \leqslant \alpha_1 \leqslant \leqslant \ldots \leqslant \alpha_m$, а каждый кластер из C_j представляет собой объединение кластеров из C_{j-1}. Подобные процедуры будем называть схемами иерархической кластеризации (СИК).

Далее Джонсон показывает, что каждая СИК приводит к специальному виду метрики между объектами и, наоборот, СИК может быть получена на основе этой метрики. Таким образом, СИК может быть исследована с помощью изучения соответствующей метрики.

Пусть в СИК даны $C_0, C_1, \ldots, C_m$ со значениями $\alpha_0, \alpha_1, \ldots \alpha_m$; определим меры расстояния $d(X_p, X_q)$ как

$$d(X_p, X_q) = \alpha_i \,, \tag{4.1}$$

где i — наименьшее целое из множества $\{0, 1, \ldots, m\}$ такое, что $X_p \in C_i$ и $X_q \in C_i$. Например, из рис. 3 видим, что $d(X_2, X_5) = 0{,}4$ и $d(X_1, X_6) = 0{,}3$. Матрица расстояний, соответствующая этой мере, будет следующей:

$$D = \begin{pmatrix} 0 & 0{,}5 & 0{,}1 & 0{,}5 & 0{,}5 & 0{,}3 \\ 0{,}5 & 0 & 0{,}5 & 0{,}4 & 0{,}4 & 0{,}5 \\ 0{,}1 & 0{,}5 & 0 & 0{,}5 & 0{,}5 & 0{,}3 \\ 0{,}5 & 0{,}4 & 0{,}5 & 0 & 0{,}2 & 0{,}5 \\ 0{,}5 & 0{,}4 & 0{,}5 & 0{,}2 & 0 & 0{,}5 \\ 0{,}3 & 0{,}5 & 0{,}3 & 0{,}5 & 0{,}5 & 0 \end{pmatrix} . \tag{4.2}$$

Можно показать, что мера расстояния (4.1) является «действенной» (bona fide) метрикой (см. определение

1.1). Наибольший интерес представляет проверка выполнимости неравенства треугольника. Пусть X, Y и Z — любые три объекта и $d(X, Y) = \alpha_j$, а $d(Y, Z) = \alpha_k$. Отсюда следует, что X и Y принадлежат некоторым кластерам, содержащимся в C_j, а Y и Z — некоторым кластерам из C_k. Однако кластер, содержащийся в C_i, где $i = \max(j, k)$, содержит и другой кластер, что следует из свойств СИК. Таким образом, X, Y и Z принадлежат одному кластеру из C_i. Далее

$$d(X, Y) \leqslant \alpha_i = \max(\alpha_j, \alpha_k), \quad d(Y, Z) \leqslant \alpha_i = \max(\alpha_j, \alpha_k)$$

и

$$d(X, Z) \leqslant \max[d(X, Y), \; d(Y, Z)]. \tag{4.3}$$

Неравенство (4.3) называется ультраметрическим неравенством. Поскольку $\max[d(X, Y), d(Y, Z)] \leqslant d(X, Y) + d(Y, Z)$, неравенство (4.3) сильнее обычного неравенства треугольника, поэтому

$$d(X, Z) \leqslant d(X, Y) + d(Y, Z).$$

Каждой СИК соответствует единственная «действенная» метрика. Наоборот, имея матрицу расстояний, например (4.2), можно построить соответствующую диаграмму-дерево (рис. 3), т. е. СИК.

Неотъемлемой частью процедуры Джонсона является определение расстояния между кластерами. Два кластера, имеющие минимальное расстояние, объединяются. Джонсон пользуется в качестве межкластерного расстояния минимальным и максимальным локальным расстоянием между кластерами (определения 1.8 и 1.9). Эти меры приводят к инвариантным методам кластеризации, т. е. результаты кластеризации инвариантны относительно монотонных преобразований матрицы сходства.

Джардайн и Сибсон [179] определяют дендограмму как функцию, отображающую интервал $(0, \infty)$ в множество отношений эквивалентности на P (множество n объектов), удовлетворяющую следующим условиям:

1) каждый кластер для данного уровня h' есть объединение кластеров на уровне h, где $0 \leqslant h \leqslant h'$;

2) для достаточно больших h все объекты объединены в один кластер;

3) для данного h существуют $\delta > 0$, такое, что множества кластеров для h и $h + \delta$ совпадают.

Условия (1), (2), (3) аналогичны соответствующим условиям Джонсона. Однако в определениях Джардай-

на и Сибсона не требуется, чтобы все объекты на уровне $h=0$ были различны. h в определениях Джардайна и Сибсона соответствует α в определении Джонсона, у которого предполагается, что $\alpha_0=h=0$. Джардайн и Сибсон обсуждают также ультраметрическое неравенство и связывают свои результаты с различными методами кластеризации. Аналогичные вопросы рассматриваются в [175], [176], [178] и [180]. В некоторых из них обсуждается также аксиоматический подход к кластерному анализу.

Гауер и Росс [139] вводят понятие минимального дерева и рассматривают его связь с односвязным кластерным анализом (минимальное локальное расстояние между кластерами, определение 1.8).

Определение 4.1. Пусть даны n точек в E_p; деревом, натянутым на данные точки, называется множество отрезков прямых (ребер), связывающих попарно эти точки таким образом, что

1) не существует замкнутых контуров;

2) через каждую точку проходит по крайней мере одна прямая;

3) дерево связано, т. е. любые две точки соединены прямой. Идеи этого определения идут из теории графов (см. [154] или [278]). Длиной дерева называется сумма длин отрезков прямых, составляющих дерево. Минимальным деревом называется дерево минимальной длины. Гауер и Росс предлагают два алгоритма для отыскания минимального дерева; там же рассматривается алгоритм, описанный в [299].

Пусть $\{d_1, d_2, \ldots, d_h\}$ обозначает множество длин ребер минимального дерева; число ребер равно h. На основе множества d_i можно построить дендограмму, группируя такие две точки, которым соответствует минимальное ребро; далее процедура совпадает с методом кластеризации по минимальному локальному расстоянию. Рис. 4 и 5 иллюстрируют эту процедуру. Описанный метод кластеризации по минимальному локальному расстоянию на минимальном дереве эквивалентен методу кластеризации на матрице расстояний, при этом элементы матрицы, которые не соответствуют ребрам дерева, во внимание не принимаются. Отсюда следует, что кластеризация на минимальном дереве эквивалентна кластеризации на матрице расстояний (Джонсон [186]. Зан [408] изучает свойства минимального дерева и возможные приложения этого понятия, в том числе к методам ступенчатой кластеризации.

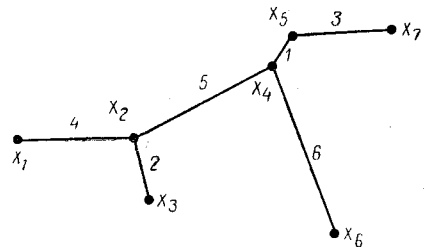


Рис. 4. Минимальное дерево для семи точек

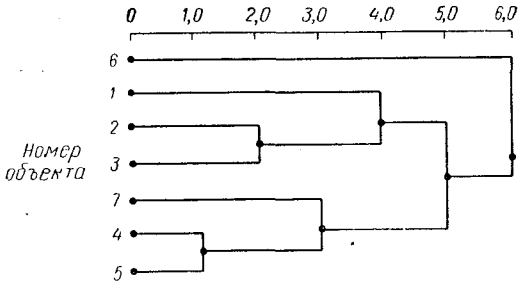


Рис. 5. Дендограмма минимального дерева из рис. 4

4.2. Сравнения дендограмм и матриц сходства

Мы уже отмечали, что в некоторых случаях матрица расстояний содержит всю информацию о соответствующей дендограмме, и наоборот (например, матрица (4.2) и дендограмма на рис. 3). Однако подобная идеальная ситуация не всегда встречается на практике. Поэтому было бы целесообразным иметь объективный способ определения, насколько хорошо дендограмма представляет свою метрику сходства или расстояния. В работе Сокала и Рольфа [335] предлагается мера соответствия между матрицей сходства и ее аппроксимацией, полученной из дендограммы. Они также предлагают метод сравнения двух дендограмм с помощью обычного коэффициента корреляции между множествами значений, которые получаются на основе этих дендограмм. Это в свою очередь приводит к методу сравнения процедур кластеризации.

Поскольку дерево или дендограмма содержит не всю информацию о матрице сходства, мы сталкиваемся с проблемой определения такого дерева, которое содержит максимум информации о матрице сходства. Таким образом, перед нами стоит проблема построения такого дерева, которое «наилучшим образом подгоняется» под данную матрицу сходства. Эта проблема была решена Хартигеном [162]. В следующих двух параграфах приводятся его основные результаты.

4.3. Основные определения

Дерево, которое обозначим τ, будем рассматривать в качестве структуры ступенчатой кластеризации. На этом мы останавливались в предыдущих параграфах. Под *узлом* (node) (вершиной графа, дерева) будем понимать либо один-единственный объект, либо кластер объектов, либо кластер кластеров. На рис. 3 каждая точка ветви представляет собой узел.

Определение 4.2. Матрица сходства S имеет *точную структуру дерева,* если $s(X_i, X_j) \leqslant s(X_p, X_q)$ всякий раз, когда узел, в котором X_i и X_j объединяются впервые, находится левее узла, в котором объединяются X_p и X_q.

Если X_p и X_q более сходны друг с другом, чем X_i и X_j, то естественно, что X_p и X_q объединяется в узел (кластер) раньше, чем X_i и X_j, или, что то же, $s(X_i, X_j) \leqslant s(X_p, X_q)$, если наименьший кластер, содержащий X_i и X_j, содержит также X_p и X_q.

Определение дерева по Джонсону [186 (параграф 2.1)] включает ультраметрическое неравенство и удовлетворяет понятию точной структуры дерева. Это же относится и к Джардайну и Сибсону [179].

Если матрица сходства S имеет точную структуру дерева, то такую же структуру будет иметь любая матрица, элементы, которой получены с помощью монотонно возрастающей функции от элементов матрицы S. Для описания матрицы S необходимо задать $n(n-1)/2$ значений, однако если S имеет точную структуру дерева, то для этого необходимо только $2n-1$ значений, которые соответствуют узлам (точкам ответвления дерева). Значение, соответствующее узлу, в котором кластеризуется пара (i, j), равно $s(i, j) = s_{ij}$.

Определение 4.3. Расстояние между двумя матрицами сходства S_1 и S_2 определим как

$$\rho(S_1, S_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} W(i, j) \{s_1(i, j) - s_2(i, j)\}^2/2,$$

где $W(i, j)$ — весовая функция сходства $s(i, j)$.

Определение 4.4. Расстоянием между матрицей сходства S и деревом τ называется

$$\rho(S, \tau) = \min_{S^*} \rho(S, S^*),$$

где S^* — любая матрица сходства с точной структурой дерева τ.

Расстоянием $\rho(S, \tau)$ можно пользоваться для определения, насколько хорошо дерево τ представляет свою матрицу сходства S. Основная проблема заключается в том, чтобы найти семейство деревьев $\{\tau_j\}$, где j изменяется от $n+1$ до $2n-1$, такое, что $\rho(S, \tau_j)$ минимально среди деревьев с j узлами. Поставленную задачу для случая, когда n не очень мало, можно решить методами дискретного программирования [292]. Однако можно найти деревья, которые будут локально оптимальными в том смысле, что ни одно дерево из семейства $\{\tau_j\}$ не может быть улучшено выполнением «локальных операций», с помощью которых деревья можно слегка изменять.

4.4. Деревья

Дерево τ определим как $\tau = [a_0, A, T]$, где a_0 обозначает корень, A — множество (конечное) узлов, включая a_0, T — отображение A в себя, такое, что для любого $k \geqslant 1$ $T^k a = a$ тогда и только тогда, когда $a = a_0$. Для данной дендограммы a_0 соответствует кластеру (узлу), содержащему все n объектов. Для данного узла b отображение T устанавливает, в какой узел переходит b после преобразования (группировки); это отображение тесно связано с данной процедурой кластеризации. Узлы $T^{-1} b$ объединяются и образуют новый кластер (узел b) и называются *семейством* b, а b называется *образующим* (parent) узлов $T^{-1} b$. Узел b называется *тривиальным,* если $T^{-1} b$ состоит из одного узла. Узел b называется *начальным* (barren), если множество $T^{-1} b$ пусто. Множество начальных узлов обозначим буквой B. Число начальных узлов обозначим $n(B)$, а общее число узлов — $n(A)$. (Очевидно, $n(B) = n$, а $n(A)$ принимают целые

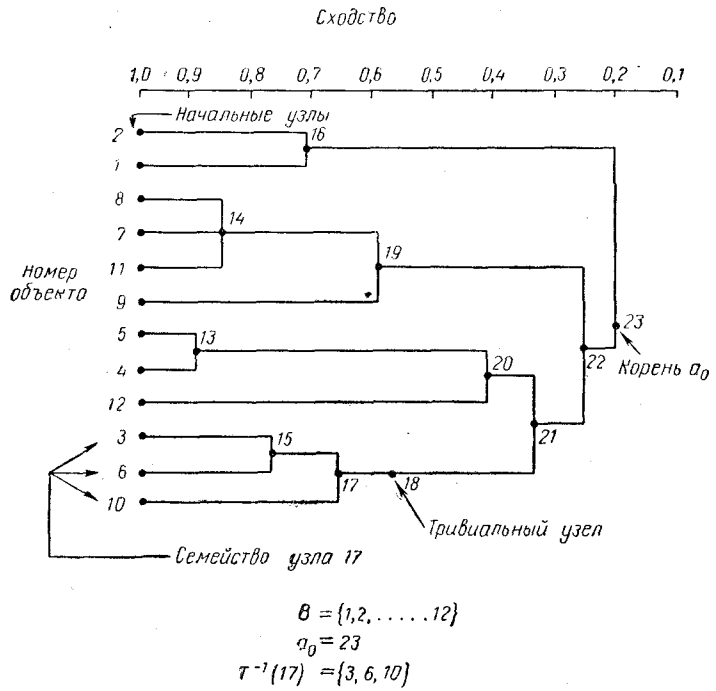


Рис. 6

значения от $n+1$ до $2n-1$.) Отображение T, которое задает дерево, начинается с начальных узлов и далее происходит в направлении корня. Таким образом, мы говорим, что основанием дерева τ является B. Эти идеи идут из теории графов [154], [278]. Введенные понятия иллюстрированы на рис. 6.

Наше изложение можно связать с соответствующим изложением этих вопросов у Джонсона, Джардайна и Сибсона, рассматривая уровень, на котором происходит образование групп. При этом каждому узлу отвечает определенный уровень расстояния или сходства α.

В обычных схемах ступенчатой кластеризации дерево не имеет узлов, подобных 14 или 18 на рис. 6. Такие узлы можно рассматривать как результат «локальных операций», которые выполняются с целью модификации данного дерева. Подобные операции будут обсуждаться в следующем параграфе.

Если два узла таковы, что для $k \geqslant 0$ $T^k a = b$, то будем писать $a \leqslant b$. Отношение $\leqslant$ задает на A частичное упорядочение. Если $a, b \in A$, то существует элемент $c = \sup(a, b)$, такой, что $a, b \leqslant c$, и из $a, b \leqslant c^*$ следует $c \leqslant c^*$, другими словами, c есть первый узел, который содержит a и b, т. е. $c = T^k a$, $c = T^m b$, а $k+m$ минимально. В терминах этих понятий дерево τ может быть определено как частичное упорядочение A, которое имеет единственный максимальный элемент (a_0) и для которого множество $\{b \mid a \leqslant b\}$ линейно упорядочено для всех a.

Дерево сходства, которое обозначим как (τ, σ), состоит из дерева и вещественнозначной функции σ на A, такой, что

$$\sigma(a) \leqslant \sigma(b) \text{ как только } b \leqslant a. \tag{4.4}$$

Действительное число $\sigma(a)$ называется сходством узла a. Любое дерево сходства может быть представлено дендограммой. В дендограмме узлам одного семейства присваивается порядок, соответствующий их узлам сходства, т. е. σ значениям; крайние значения задаются произвольно. С помощью этой процедуры, которая начинается с корня, в начальных узлах можно установить линейный порядок. Начальные узлы располагаются вертикально в порядке вычислений и горизонтально соответственно значениям своего сходства, узлы с большим сходством располагаются левее. Другие узлы располагаются вертикально в центрах соответствующих семейств, которые также располагаются вертикально и горизонтально соответственно значениям сходства. На основе дендограммы может быть построено полное дерево сходства τ. Действительно, дендограмма есть не что иное, как графическое представление абстрактного дерева $\tau = [a_0, A, T]$.

Матрица сходства S имеет *точную структуру дерева τ на B*, т. е. $S \in \tau$, если для некоторого σ, (τ, σ) есть дерево сходства

$$s(i, j) = \sigma(\sup_{\tau}(i, j)) \tag{4.5}$$

для всех $i, j \in B$. Другими словами, величина сходства между двумя узлами $i, j \in B$ равна сходству первого узла, в котором они кластеризуются с помощью T-отображения. Это определение аналогично определению Джонсона [186] и Джардайна и Сибсона [179]; рас-

стояние между двумя узлами i и j полагается равным тому уровню расстояния, на котором происходит объединение i и j. Если матрица сходства S имеет точную структуру дерева, то $[n(B)]^2$ (т. е. $n(n-1)/2$) элементов матрицы S могут быть определены по $n(A)$ значениям σ.

Расстояние между двумя матрицами сходства S_1 и S_2 определяется как

$$\rho(S_1, S_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} W(i,j)\,[s_1(i,j) - s_2(i,j)]^2/2, \quad (4.6)$$

где W — симметричная весовая функция.

Расстояние между матрицей S и деревом τ определяется как

$$\rho(S, \tau) = \inf_{S^* \in \tau} \rho(S, S^*), \quad (4.7)$$

где $S^* \in \tau$ означает, что S^* имеет точную структуру дерева. Матрица сходства S, которая минимизирует $\rho(S, S^*)$, называется *приближением S по τ*. Итак, $s^*(i,j)$ есть сходство первого узла, в котором i и j группируются впервые, т. е. $s^*(i,j) = \sigma(\sup_{\tau}(i,j))$, где σ — некоторая вещественнозначная функция на A, такая, что $\sigma(a) \geqslant \geqslant \sigma(b)$, если $a \leqslant b$. Отсюда следует, что

$$\rho(S, \tau) = \inf \sum_{i=1}^{n} \sum_{j=1}^{n} W(i,j)\,[s(i,j) - \sigma(\sup_{\tau}(i,j))]^2/2,$$

где σ может быть выбрано произвольно и единственным ограничением является выполнение определенных неравенств. Такое определение $\rho(S, \tau)$ приводит нас к задаче квадратичного программирования; при этом существует единственное решение, которое и приводит Томпсон [358]. Однако если единственным условием минимизации $\rho(S, S^*)$ является $S^*(i,j) = \sigma(\sup_{\tau}(i,j))$ для произвольного σ, то

$$\sigma(c) = \sum_{\substack{\sup(i,j)=c \\ \tau}} W(i,j)\,s(i,j) \Big/ \sum_{\substack{\sup(i,j)=c \\ \tau}} W(i,j). \quad (4.8)$$

Если эта функция удовлетворяет неравенству $\sigma(a) \geqslant \sigma(b)$ для $a \leqslant b$, то S^* является приближением S по дереву τ. В действительности наибольший интерес представляют такие деревья τ, для которых это неравенство выполняется (см. [162]).

Нахождение приближения S по τ эквивалентно отысканию таких деревьев τ, для которых величина $\rho(S, \tau)$ принимает минимальные значения. Определение расстояния $\rho(S, \tau)$ позволяет нам выбрать «оптимальное» представление S деревом. Расстояние $\rho(S, \tau)$ есть *средний квадрат ошибки* при подстановке вместо S матрицы близости с точной структурой дерева τ.

Для любого дерева τ на B обобщение W, S на $A \times A$ производится следующим образом:

$$W(a,b) = \sum_{i \leqslant a} \sum_{j \leqslant b} W(i,j); \quad (4.9)$$

$$W(a,b)\,s(a,b) = \sum_{i \leqslant a} \sum_{j \leqslant b} W(i,j)\,s(i,j). \quad (4.10)$$

Веса узлов приближения и их коэффициенты сходства определяются следующим образом:

$$w(c) = W(c,c) - \sum_{Tc' = c} W(c', c') \quad (4.11)$$

и

$$w(c)\,\sigma(c) = W(c,c)\,s(c,c) - \sum_{Tc' = c} W(c', c')\,s(c', c'). \quad (4.12)$$

σ, определенные из уравнений (4.8) и (4.12), совпадают, поскольку

$$W(c,c) = \sum_{i \leqslant c,\ j \leqslant c} W(i,j) = \sum_{c' \leqslant c} \sum_{\substack{\sup(i,j)=c \\ \tau}} W(i,j).$$

Поэтому

$$w(c) = W(c,c) - \sum_{Tc' = c} W(c', c') = \sum_{\substack{\sup(i,j)=c \\ \tau}} W(i,j).$$

Таким же образом

$$w(c)\,\sigma(c) = \sum_{\substack{\sup(i,j)=c \\ \tau}} W(i,j)\,s(i,j).$$

Теперь средний квадрат ошибки $\rho(S, \tau)$ может быть записан как

$$\rho = (S, \tau) = \sum_{i,j \in B} W(i,j)\,s^2(i,j) - \sum_{a \in A} w(a)\,\sigma^2(a), \quad (4.13)$$

поскольку

$$\rho(S, \tau) = \sum_{i,j \in B} W(i,j)\,[s(i,j) - \sigma(\sup_{\tau}(i,j))]^2 =$$

$$= \sum_{c \in A} \sum_{\substack{\sup(i,j)=c \\ \tau}} W(i,j)\,[s(i,j) - \sigma(c)]^2 =$$

$$= \sum_{c\in A} \sum_{\sup(i,j)=c} [W(i,j)s^2(i,j) - W(i,j)\sigma^2(c)] =$$

$$= \sum_{i,j\in B} W(i,j)s^2(i,j) - \sum_{c\in A} w(c)\sigma^2(c).$$

Уравнения (4.9)—(4.13) применяются для приближения дерева к матрице сходства S. Для данного S цель заключается в том, чтобы найти такие деревья τ, для которых $\rho(S,\tau)$ минимально, т. е. найти такие коэффициенты сходства узлов приближения $\sigma(c)$, которые минимизируют взвешенную сумму квадратов (среднее квадрата ошибки):

$$\sum_{c\in A} \sum_{\substack{\sup(i,j)=c \\ \tau}} W(i,j)\,(s(i,j)-\sigma(c))^2 .$$

4.5. Локальные операции на деревьях

Основная цель приближения τ к S заключается в том, чтобы найти такие деревья τ, которые обращают $\rho(S,\tau)$ в минимум. Однако всегда существует такое дерево τ_{j+1}, что $\rho(S,\tau_{j+1})\leqslant\rho(S,\tau_j)$. Поэтому целью является отыскание семейства деревьев $\{\tau_j,\ n(B)<j<2n(B)+1\}$, таких, что $\rho(S,\tau_j)$ дает минимум для каждого j. Единственный существующий метод отыскания оптимального семейства состоит в полном переборе всех возможных деревьев на B и вычислении для каждого дерева соответствующего значения $\rho(S,\tau)$. Однако этот метод практически неосуществим из-за того, что число деревьев очень быстро растет с ростом $n(B)$. Таким образом, мы приходим к понятию «локально оптимального» семейства. Процедура начинается с семейства $\{\tau_j\}$, над которым производятся некоторые итеративные преобразования, которые составляют так называемое множество «локальных операций» L. Локальной операцией $L\in$L называется любая операция, которая изменяет дерево τ и в результате которой получается другое дерево L_τ. Семейство деревьев $\{\tau_j\}$ называется L-оптимальным, если для каждого $\tau_j\,\rho(S,\tau_{i_L})\leqslant\rho(S,L\tau_j)$, где j_L обозначает число узлов в $L\tau_j$. Это означает, что семейство не может быть улучшено с помощью операций L. При оперировании на деревьях семейство $\{\tau_j\}$ итеративно преобразуется в L-оптимальное семейство. Семейство деревьев $\{\tau_j\}$ назовем локально оптимальным на множестве локальных операций L, если это семейство является L-оптимальным для каждого $L\in$L. Желаемыми свойствами локальных операций являются: 1) легкость вычислений, 2) возможность ρ-сокращений. Оценить операцию в терминах свойства (2) довольно трудно, за исключением некоторых экспериментальных способов, поэтому свойство (1) является решающим при выборе того или иного семейства операций.

Можно показать, что если коэффициенты сходства узлов приближения, вычисленных по формуле (4.8), не удовлетворяют неравенству $\sigma(a)\geqslant\sigma(b)$ для любых $a\leqslant b$, то можно найти дерево с количеством узлов, меньших на единицу, и которое приближает S так же хорошо, т. е. оба дерева приводят к одному и тому же значению ρ. Это означает, что поиск локально оптимальных деревьев можно ограничить классом деревьев сходства. Аргументацию этого предложения читатель найдет в работе Хартигена [162, параграф 5].

Локальная операция L на τ приводит к новому дереву $L\tau$ с подогнанными весами и коэффициентами сходств узлов, которые обозначим соответственно Lw и $L\sigma$. Получаемое улучшение (если таковое будет) представляет собой разность $\rho(S,L\tau)-\rho(S,\tau)$. Значение $\rho(S,L\tau)$ может быть вычислено на основе Lw и $L\sigma$. Если через $m_1(L)$ обозначить число ω и σ, изменяющихся $L\tau$ при операции L, то $m_1(L)$ можно использовать как показатель «трудоемкости» вычисления $\rho(S,L\tau)$.

После того как локальная операция выполнена, основная корректировка заключается в изменении некоторых значений расширенных матриц W, S. Все другие величины, связанные с деревом, должны быть изменены одинаковым образом. Процедура заключается в том, чтобы оценить данную серию операций на дереве τ и выполнить ту, которая приводит к наибольшему уменьшению ρ. Результат локальной операции L, расстояние $\rho(S,L\tau)$, сравнивается с $\rho(S,\tau^*)$ или $\rho(S,S^*)$, где τ^* — оптимальное дерево с j_L узлами, S^* — соответствующая матрица сходства. Следующей вычислительной задачей является определение матрицы сходства S^* с точной структурой дерева τ^*, для которой $\rho(S,S^*)$ минимально.

Три локальные операции, выполняемые на деревьях:

1) операция разветвления. Операция разветвления из a в b обозначается $L(a,b)$; это операция, которая соединяет некоторый узел a из семейства Ta и некоторый другой узел дерева b, причем $b\not< a$. Функция T из-

меняется на функцию LT так, что $(LT)a'=Ta'$ для $a'\neq a$ и $(LT)a=b$. Операция разветвления не меняет числа узлов дерева;

2) операция исключения $K(a)$ исключает узел a из дерева и переводит семейство a в $T(a)$. Эта операция определена для любых узлов за исключением корня и начальных узлов. В терминах отображения T это означает $(KT)a'=Ta'$ для $a'\neq a$, $Ta'\neq a$ и $(KT)a'=Ta$ для $Ta'=a$;

3) операция включения $M(a)$ включает новый узел a^* в дерево τ между a и Ta. Если a — корень, то a^* — новый корень. В терминах T имеем: $(MT)a'=Ta'$ для $a'\neq a, a^*$; $(MT)a=a^*$; $(MT)a^*=Ta$ для $a\neq a_0$ и $(MT)a^*=a^*$ для $a=a_0$.

Операция включения и исключения изменяет число узлов дерева τ.

Иллюстрации этих операций читатель найдет в статье Хартигена. Он также дает примеры применения этой процедуры для отыскания локально оптимального дерева и показывает, как F-критерий может быть использован для определения приемлемого размера дерева. Там же рассматриваются другие определения точной структуры дерева и другие, отличные от деревьев, структуры сходства.

ГЛАВА 5

КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ ОЦЕНИВАНИЯ ФУНКЦИИ ПЛОТНОСТИ

5.1. Модальный анализ

Один. из методов кластеризации, рассмотренных в главе 1, — метод по минимальному локальному расстоянию — приводит к удлинению кластеров. Этот метод не принадлежит к классу методов с минимальной дисперсией, которые рассматривались в параграфе 1.6.

Применение для кластеризации метода минимального локального расстояния благодаря цепной тенденции может привести к ряду «плотных» кластеров, которые перемежаются, соединяются «редкими, неплотными» кластерами. В случае одной характеристики гистограмма будет иметь вид мультимодального распределения. Желательно пользоваться методами, которые бы определяли моды этого распределения и соответствующие им отдельные кластеры.

Для массивов среднего объема Уишарт [396] предложил метод кластеризации, который он назвал *модальным анализом*. Этот метод им же был обобщен на случай большого числа наблюдений. Его процедура начинается с выяснения вопроса о мультимодальности данных. В случае одной характеристики необходимо построить гистограмму и вычеркнуть данные с малой частотой (седловые области). Тогда соответствующий кластер можно установить для каждой модальной области. Данные, принадлежащие седловой области, относятся к ближайшей моде. В случае p-характеристик этот метод становится неудобным. Если каждая ось измерения разбита на k классов, то мы получим p^k p-мерных прямоугольников. Для определения, в какой из классов необходимо отнести то или иное наблюдение,

требуется ответить на ряд вопросов, одним из которых является выбор оси измерения. Эту проблему можно обойти, если пользоваться сферическими областями. Одноуровневый алгоритм Уишарта может быть записан следующим образом:

а) выбираем значения порогового расстояния r и пороговой частоты f;

б) вычисляем матрицу сходства S;

в) для каждой точки находим частоту попадания точек f_i, лежащих на расстоянии меньшем r;

г) точку с частотой меньшей чем f_i удаляем;

д) кластеризуем оставшиеся точки концентрации по методу минимального локального расстояния;

е) распределяем точки, исключенные на шаге (г), по кластерам, полученным в (д), соответственно некоторому критерию. (Например, каждую точку, не являющуюся точкой концентрации, приписываем к кластеру, для которого расстояние между данной точкой и соответствующей точкой концентрации минимально.)

Далее, Уишарт предложил ступенчатый алгоритм, который выполнял только задачу модального анализа. Для этого необходимо было задать лишь пороговое значение для частоты f. На первом и последнем цикле алгоритма определяется один кластер. На некотором промежуточном цикле получается максимальное число кластеров. Для унимодальных данных анализ приводит к одному кластеру. За полным описанием этого ступенчатого алгоритма отсылаем читателя к работе [396].

Заде [407] ввел понятие «размытого» множества, его процедура имеет много общего с модальным анализом Уишарта. По Заде, если E — пространство точек, размытое множество A в E характеризуется функцией семейства (характеристической функцией) $f_A(x)$, которая каждой точке в E ставит в соответствие число в интервале [0, 1], причем величина $f_A(x)$ представляет собой степень «принадлежности» x к множеству A. Смысл $f_A(x)$ аналогичен пороговому значению f Уишарта.

Гитман и Левин [129] предлагают алгоритм, который разбивает выборку из мультимодального размытого множества на унимодальные размытые множества; а затем применяется алгоритм кластеризации многомерных наблюдений, в результате которого получаются однородные группы. Эта процедура также аналогична модальному анализу Уишарта.

Результат кластеризации, основанный на модальном анализе, сильно зависит от оценивания положения моды, поэтому различные методы оценивания мультимодальных многомерных функций плотностей приведут и к новым методам кластеризации. Этой проблеме мы и посвятим оставшуюся часть этой главы.

5.2. Оценивание функции плотности вероятности

По вопросу оценивания функции плотности вероятности имеется много работ. Мы не ставим себе цели дать обзор методов оценивания функций плотности. За коротким обзором существующих методов отсылаем читателя к Брайену [40]. Брайен предлагает один метод оценивания многомерных функций плотности вероятности, который называет методом ядра; он также строит метод кластеризации, который основывается на оценке функции плотности.

Метод ядра оценивания функции плотности связан с линейным интегральным преобразованием

$$G(x) = \int K(x-y)\, g(y)\, dy.$$

Это преобразование устанавливает соответствие между функциями $G(x)$ и $g(y)$. Функция $K(x—y)$ называется *ядром* преобразования (см. [209]). Метод ядра также иногда называют *оцениванием с взвешенным средним* (см. [95], [283], [298]).

В методе ядра функция плотности вероятности $f(x)$ оценивается по формуле

$$\hat{f}(x) = \int K(x-u)\, dF_n(u) = \frac{1}{n} \sum_{j=1}^{n} K(x-x_j),$$

где $K(x—y)$ — ядро, а $F_n(u)$ — эмпирическая функция распределения. Розенблат [298] предложил само ядро рассматривать как некоторую функцию плотности, т. е.

$$K(x) \geqslant 0, \quad \int K(x)\, dx = 1.$$

Сейчас мы рассмотрим основные моменты процедуры оценивания плотности вероятности, предложенной Брайеном [40].

Пусть $X_1, X_2, \ldots, X_n$ обозначает случайную выборку объема n из некоторой функции плотности $f(x)$, имеющей невырожденную матрицу ковариаций Σ. С помо-

щью рассматриваемого метода, который аналогичен методу Какоулоса [42], оценивается функция плотности в виде:

$$\hat{f}(x) = \frac{1}{n\alpha^p} \sum_{j=1}^{n} K\left(\frac{x-x_j}{\alpha}\right),$$

где K — ядро. В качестве ядра выбирается функция плотности многомерного нормального распределения с математическим ожиданием, равным нулю, а ковариационной матрицей S, т. е.

$$K(x) = \frac{1}{(2\pi)^{p/2}|S|^{1/2}} e^{-(1/2)x^T S^{-1} x}. \qquad (5.1)$$

Оценкой тогда будет:

$$\hat{f}(x) = \frac{1}{n\alpha^p (2\pi)^{p/2}|S|^{1/2}} \sum_{j=1}^{n} e^{-\frac{1}{2\alpha^2}(x-x_j)^T S^{-1}(x-x_j)}, \qquad (5.2)$$

где $X_1, X_2, \ldots, X_n$ обозначают векторы наблюдений, а $S = (1/n)\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T$ — выборочная матрица ковариаций, которая предполагается невырожденной, откуда следует, что $n>p$ и S положительно определена.

Легко показать, что $\hat{f}(x)$ из (5.2) является функцией плотности вероятности, другими словами,

$$\hat{f}(x) \geqslant 0 \quad \text{и} \quad \int \hat{f}(x)\,dx = 1.$$

Квадратичная форма $x^T S^{-1} x$ в экспоненте $K(x)$ в (5.1) является расстоянием Махаланобиса между x и 0. Вместо матрицы S можно привлечь матрицу I*; это приведет к квадрату евклидова расстояния между x и 0, т. е. $d_2^2(x, 0) = x^T x$. Выбор S и I эквивалентен выбору между расстоянием Махаланобиса и евклидовым расстоянием.

Евклидово расстояние проще и легче вычисляется. Однако расстояние Махаланобиса имеет много преимуществ. Например, как показано в параграфе 1,3, это расстояние инвариантно по отношению ко всем невырожденным преобразованиям**. Это означает, что $\hat{f}(x)$ совпадает с $\hat{f}(Ax)$, где A — невырождено. Таким образом, выбор масштаба не влияет на $\hat{f}(x)$. Это перестает быть верным для евклидова расстояния.

Другим свойством расстояния Махаланобиса является то, что оно делает некоторые критерии кластеризации эквивалентными. Следующие три критерия кластеризации при пользовании расстоянием Махаланобиса эквивалентны: 1) tr W; 2) $|T|/|W|$; 3) tr $W^{-1}B$, где T, B и W — соответственно матрицы полного, межгруппового и внутригруппового расстояния, которые обсуждались в параграфе 1.4. Критерии (2) и (3) были предложены Фридманом и Рубиным [122], ими же были исследованы свойства этих критериев. В следующем параграфе мы рассмотрим процедуру кластеризации, которая приводит к оптимальному, по меньшей мере в локальном смысле, разбиению на группы относительно критериев (2) и (3). Метод, который был рассмотрен в параграфе 1.5, приводит к оптимальному разбиению с точки зрения критерия (1).

Одной из проблем применения оценки $\hat{f}(x)$ в (5.2) является выбор значения α. Выбор того или иного значения очень важен, и в случае плохого выбора оценка будет неудовлетворительной.

Выбор α основан на неравенстве теории информации:

$$r = E\left\{\ln\frac{f(x)}{\hat{f}(x)}\right\} = \int f(x)\ln\frac{f(x)}{\hat{f}(x)}\,dx \geqslant 0. \qquad (5.3)$$

Кульбак [217] доказал, что равенство в (5.3) достигается тогда и только тогда, когда $f(x) = \hat{f}(x)$ для почти всех x. Процедура заключается в том, чтобы минимизировать r. Неравенство (5.3) можно переписать следующим образом:

$$m = E[\ln\hat{f}(x)] = \int f(x)\ln\hat{f}(x)\,dx \leqslant \int f(x)\ln f(x)\,dx = E[\ln f(x)],$$

причем равенство достигается в том и только в том случае, когда $f(x) = \hat{f}(x)$ для почти всех x. Как видим, минимизация r эквивалентна максимизации m. Процедура выбора α заключается в максимизации не самого зна-

* I — единичная матрица порядка $n \times n$.— *Примеч. пер.*

** Линейным.— *Примеч. пер.*

чения m, а его оценки. Если $\hat{f}$ построено на $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_k$ — другая выборка объема k, то оценка m равна:

$$\hat{m}=\frac{1}{k}\sum_{i=1}^{k}\ln\hat{f}(y_i).$$

Если второй выборки не существует, то оценку для $\hat{f}$ строят на основе $x_1, x_2, \ldots, x_n$; в этом случае она равна:

$$\hat{m}=\frac{1}{k}\sum_{i=1}^{k}\ln\hat{f}(x_i). \qquad (5.4)$$

Однако эта оценка будет смещена и может привести к отрицательным значениям α. Смещение оценки (5.4) может быть уменьшено методом *складного ножа* (jackknifing) [81], [141]. Пусть $\hat{h}_j(x)$ — оценка $f(x)$ (x_j опущен). Тогда

$$\hat{h}_j(x)=\frac{1}{(n-1)\alpha^p}\sum_{\substack{i=1\\ i\neq j}}^{n}K\left(\frac{x-x_i}{\alpha}\right),$$

а $\hat{m}$ определяется равенством:

$$\hat{m}=\frac{1}{n}\sum_{j=1}^{n}\ln\hat{h}_j(x_j).$$

Производная $\hat{m}$ после преобразований равна:

$$\frac{d\hat{m}}{d\alpha}=\frac{1}{n\alpha^3}\sum_{j=1}^{n}\frac{\sum_{i\neq j}D^2(x_i,x_j)\,e^{-\frac{1}{2\alpha^2}D^2(x_i,x_j)}}{\sum_{i\neq j}e^{-\frac{1}{2\alpha^2}D^2(x_i,x_j)}}-\frac{p}{\alpha},$$

где

$$D^2(x_i,x_j)=(x_i-x_j)^TS^{-1}(x_i-x_j).$$

Оптимальное значение α находят из уравнения $\frac{d\hat{m}}{d\alpha}=0$. Для решения этого уравнения могут быть применены различные методы, например метод Ньютона—Рафсона. Эти методы рассматриваются в главе 3, Изаксона и Келлера [172].

5.3. Кластеризация на основе оценивания функции плотности

В предыдущем параграфе было указано, что одним из преимуществ применения оценки плотности $\hat{f}(x)$ (5.1) является то, что при этом три критерия группировки (1) W; (2) $|T|/|W|$ и (3) $\mathrm{tr}\,W^{-1}B$ становятся эквивалентными. Для этого векторы наблюдений $y_1, y_2, \ldots, y_n$ необходимо заменить на $x_i=y_i-\bar{y}$, и так что новое семейство векторов будет иметь среднее, равное 0; а затем векторы x преобразовать по закону CX, где C — невырожденная матрица, такая, что $CTC^T=I$. Как следует из теоремы, доказанной в [40], евклидово расстояние в преобразованном пространстве пропорционально расстоянию Махаланобиса первоначального пространства. Отсюда можно сделать следующий вывод: алгоритмом кластеризации можно пользоваться в соответствии с одним из трех критериев без предварительного преобразования, если первоначально пользовались расстоянием Махаланобиса.

Большинство из методов кластеризации, которые обсуждались в предыдущих главах, были сформулированы на эвристической и интуитивной основе и имели детерминированную природу. Методы, основанные на оценивании функции плотности, составляют содержание статистического подхода и приводят к хорошо обоснованному понятию кластера (по крайней мере не хуже, чем в детерминированном случае). Один из основных моментов в статистическом подходе — оценивание моды. При этом имеются два способа: 1) моды оцениваются непосредственно из наблюдений; 2) сначала оценивается многомерная функция плотности $\hat{f}(x)$, на основе которой затем вычисляются моды.

При эвристических подходах выбор метода кластеризации зависит в основном от данных. При статистическом подходе кластер определяется в терминах характеристик функции плотности, которой соответствуют наблюдения. Этот подход более точен, так как при этом мы можем более точно сказать, что является целью кластерирования; статистическое оценивание даст объективный подход к достижению этой цели, а статистическая проверка поможет ответить на вопрос, как близко результат лежит от цели.

При статистическом подходе к кластеризации, предложенном в [40], в первую очередь оценивается многомерная функция плотности для произведенных наблюдений. Затем с целью распределения наблюдений по кластерам применяется метод «подъема на холм» (hill chimbing); кластеры определяются в терминах моды плотности $\hat{f}(x)$. Процедура «подъема на холм» заключается в следующем. Сначала оценивается $\hat{f}(x)$. Затем для каждого наблюдения x_i определяются два ближайших элемента (по минимальному расстоянию). Наблюдение становится частью «пути» или «холма» $\hat{f}(x)$, если ближайшее дает большее значение $\hat{f}(x)$. В результате перебора всех наблюдений получатся пути, которые поднимаются на «холмы» (или холм) $\hat{f}(x)$. Наблюдения на каждом пути группируются и образуют кластеры.

Путь заканчивается наблюдением, для которого ближайшее другое приводит к меньшему значению $\hat{f}(x)$. Алгоритм, основанный на минимальном расстоянии, может привести к большому числу путей (кластеров). Для устранения подобной ситуации для каждого наблюдения выбирается другое, расстояние до которого равно второму минимальному значению. Если для элемента (наблюдения) x_i первый ближайший элемент равен x_k, а второй x_j и $\hat{f}(x_k) \leqslant \hat{f}(x_i) \leqslant \hat{f}(x_j)$, то путь продолжается. Процедура первых двух ближайших наблюдений применяется для построения начального приближения разбиения на группы.

Два пути по обеим сторонам «холма» (моды) в этой процедуре будут всегда несвязанными. Для устранения последствий этой ситуации сравним $\hat{f}(x)$ в трех точках между каждой вершиной и ее ближайшей вершиной. Значения $\hat{f}(x)$ в этих точках указывают либо на то, что две ближайшие вершины составляют «холм» (в этом случае наблюдения, соответствующие двум путям, объединяются), либо на «долину», которая лежит между холмами, — в этом случае пути оставляем разомкнутыми. Процедура повторяется до тех пор, пока все наблюдения не будут соответствовать одному холму или пока между каждой вершиной и ближайшей к ней не будет

долины. Три точки между ближайшими вершинами могут быть выбраны, например, по следующему правилу: первая точка соответствует одной четвертой отрезка прямой, соединяющей вершины, вторая — одной второй отрезка, третья — трем четвертым отрезка.

Брайен [40] описывает вычислительную программу, которая выполняет оценивание функции плотности и производит кластеризацию по методу, изложенному выше. Он приводит примеры, для которых этот метод привел к удовлетворительным результатам. В одном из этих примеров привлекаются хорошо известные данные наблюдений Фишера за ирисами; этот пример будет описан в следующей главе.

5.4. Замечания

Из параграфов 5.2 и 5.3, очевидно, следует, что процедура Брайена может быть модифицирована многими способами. Например, могут быть применены различные оценки для функции плотности. Однако оценка Брайена в отношении величины α обладает определенной гибкостью. Решение некоторых вопросов его процедуры кластеризации довольно произвольно и может быть модифицировано, однако все это имеет второстепенное значение.

Другое определение кластера, с помощью некоторой вероятностной модели, предложено Лингом [231]. Он дает определения кластера и двух показателей, которые служат мерами компактности и относительной изоляции. Далее он развивает соответствующую вероятностную модель для выборочных распределений этих показателей.

Вольф [402] описывает кластерный метод, основанный на разбиении смеси многомерных распределений, и приводит полное описание программы для IBM 360/65 этого метода.

Кунц и Фукунага [208] приводят общее выражение для критерия при непараметрической кластеризации, которое аналогично (4.6). Они описывают процедуру кластеризации, основанную на их критерии; эта процедура отыскивает «долины», т. е. определяет области с низкими частотами (седловины).

ГЛАВА 6

ПРИЛОЖЕНИЯ

6.1. Приложение к регистрации отдаленных объектов

В исследовании проблемы регистрации отдаленных объектов [167], [226], [227] имеют дело с участком (экраном), который представляет собой прямоугольную область, площадь которой составляет r строк (линий сканирования) и c столбцов (число различаемых элементов объектов) на одной сканируемой линии*. Для каждой точки (объекта I) имеется вектор измерений (характеристик) $p \times 1$ X_{ij}, $i=1, 2, \ldots, r$; $j=1, 2, \ldots, c$. Для рекогносцировки участка необходимо «как можно эффективнее» решить $rc=n$ задач на распознавание (участок рекогносцируется последовательно точка за точкой).

Для решения этой задачи, т. е. для кластеризации наблюдений отдаленных объектов (мультиспектральных данных сканирования) центр пилотируемых космических кораблей воспользовался программой Болла и Холла [15], [16], [18] ИСОМАН (ISODATA — Iterative Self—Organizing Data Analysis Technique) (итеративный самообучающийся метод анализа наблюдений). Цель процесса кластеризации двояка [194]: а) проверить однородность мультиспектральных данных сканирования, т. е. необходимость разделения класса различаемых элементов на несколько унимодальных подклассов и б) кластеризовать данные сканируемой линии, т. е. классифицировать объекты по группам.

Итеративная процедура Болла и Холла [15], [16], [18] была коротко описана в параграфе 1.6. Сначала выбирается k кластеров, которые представляют собой k случайно выбранных точек; оставшиеся объекты приписываются к кластерам с ближайшим центром. Затем вычисляются центры кластеров и два кластера I и J объединяются, если D^2_{IJ} меньше заданного порогового значения r. Кластер расщепляется, если внутригрупповая дисперсия кластера S_x^2 по любой характеристике x превышает пороговое значение S^2. Таким образом, дисперсия S_I^2 результирующего кластера I ограничена неравенством $S_I^2 \leqslant pS^2$, где p — число характеристик. Вместо первоначальных центров подставляются новые и процесс продолжается до полной стабилизации (до полной сходимости). Описание программ ИСОМАН содержится у Холли [166]. Первоначальный вариант программы ИСОМАН основан на евклидовой метрике. Вариант программы, описанной в [166], основан на применении взвешенной евклидовой метрики.

Кан и Холли [194] предложили окончательный вариант ИСОМАН. При распределении объектов по ближайшим кластерам (центрам кластеров) вместо евклидовой метрики они воспользовались метрикой l_1*. Мера вариации для каждой характеристики S^2_x оставалась той же. Для вычисления D^2_{IJ} бралось взвешенное евклидово расстояние.

Опишем применение окончательной рекомендации программы ИСОМАН Кана и Холли для мультиспектральных данных сканирования. Пример взят из работы Кана и Холли [194]. Имеется $r=35$ линий сканирования и $c=45$ столбцов или выборочных точек на каждой линии сканирования. Таким образом, имеем $n=35 \cdot 45=1575$ наблюдений. Существуют четыре источника (способа) получения мультиспектральных данных сканирования, т. е. $p=4$. Пороговое значение расщепления равно 4,5, а пороговое значение объединения 3,2. Это означает, что на некоторой итерации кластер расщепляется по j-му измерению, если дисперсия по j-му измерению превосходит значение 4,5. Наоборот, если два кластера находятся на расстоянии меньшем, чем 3, 2 единицы, то они объединяются в один кластер.

Программе потребовалось 12 итераций для образования семи кластеров с числом элементов, равным соответственно 565, 132, 219, 201, 180, 224 и 54.

* Т. е. разрешающая способность прибора, сканирующего (регистрирующего) объекты. — *Примеч. пер.*

* См. параграф 1.3. — *Примеч. пер.*

Таблица 6.1. **Итоговые статистики для каждого кластера**

№ кластера		Источник 1	2	3	4
1	среднее	182,59	176,91	187,33	200,39
	стандартное отклонение	1,779	2,671	1,814	1,862
2	среднее	178,35	172,44	173,98	187,72
	стандартное отклонение	3,619	2,840	2,532	2,720
3	среднее	166,68	163,63	167,10	181,35
	стандартное отклонение	2,250	2,150	2,614	2,541
4	среднее	179,42	174,99	181,01	194,69
	стандартное отклонение	2,308	2,259	2,337	2,130
5	среднее	179,77	170,24	157,64	166,74
	стандартное отклонение	2,200	1,728	2,134	2,348
6	среднее	163,10	158,60	159,57	174,53
	стандартное отклонение	2,362	2,645	3,356	2,946
7	среднее	181,89	173,52	164,56	173,74
	стандартное отклонение	2,738	1,940	3,306	4,006

В табл. 6.1 приводятся итоговые статистики для каждого кластера, а табл. 6.2 представляет собой матрицу межкластерных расстояний на конец 12-й итерации. Данные для примера Кана и Холли были взяты из [409]. Остальные детали рассмотрены в [194].

Таблица 6.2. **Таблица межкластерных расстояний**

Кластер / Кластер	1	2	3	4	5	6	7
1	0,0	8,7	16,0	4,5	22,3	19,7	13,6
2	8,7	0,0	6,5	4,2	10,9	9,9	5,5
3	16,0	6,5	0,0	11,1	9,9	4,4	8,2
4	4,5	4,2	11,1	0,0	16,5	14,7	9,4
5	22,3	10,9	9,9	16,5	0,0	9,6	4,0
6	19,7	9,9	4,4	14,7	9,6	0,0	10,0
7	13,6	5,2	8,2	9,4	4,0	10,0	0,0

6.2. Применение метода оценивания функции плотности для данных Фишера по ирису [40]

Данные Фишера по ирису состоят из четырех характеристик ($p=4$); при этом рассматриваются три сорта ирисов. Для каждого сорта имеется выборка объемом 50 наблюдений. Сорта ирисов следующие: ирис «сетоса», ирис «версикалор», ирис «вирджиника»; четырьмя характеристиками служат: длина стебля, ширина стебля, длина лепестка и ширина лепестка. Кластерный анализ с помощью расстояния Махаланобиса привел к девяти кластерам, что в табл. 6.3 показано по сортам.

Таблица 6.3. **Число элементов, принадлежащих кластерам, разбитые по сортам**

Сорт / Кластер	1	2	3	4	5	6	7	8	9
Ирис «сетоса»	49	1	0	0	0	0	0	0	0
Ирис «версикалор»	0	14	13	17	3	3	0	0	0
Ирис «вирджиника»	0	7	5	6	18	0	3	5	6

Если отнести кластер 1 к ирису «сетоса», кластеры 2, 3, 4, 6 — к ирису «версикалору», а кластеры 5, 7, 8, 9 — к ирису «вирджиника», то получится 22 класса. Процесс кластеризации с помощью евклидова расстояния приводит к 8 классам.

ГЛАВА 7

ИСТОРИЧЕСКИЕ ЗАМЕЧАНИЯ

В последние годы к области классификации наблюдений, а в особенности к кластерному анализу был проявлен огромный интерес. За это время в различных журналах появилось большое число публикаций по кластерному анализу, которые охватывают самые разнообразные вопросы. В то же время отсутствуют работы, которые бы объединяли полученные результаты и в которых читатель нашел бы последовательное изложение всех вопросов. Первые шесть глав этой книги (может быть, за исключением главы 2) являются попыткой в этом направлении.

В этой главе мы сделаем некоторые замечания о развитии кластерного анализа за последние четыре десятилетия.

Первоначальное описание и определение предмета, известного сейчас под названием «кластерный анализ», было сделано Трионом [361] в 1939 г. В недавно вышедшей книге Триона и Бейли [371, 1970] они рассматривают вычислительную систему ВСПУ, предназначенную для решения задач кластеризации и факторного анализа в области социологии. В книге Фишера [108, 1968] рассматриваются специальные методы, применяемые в задачах агрегирования в экономике. Коул [57, 1969] рассматривает работы, представленные на коллоквиум по численной таксономии. В недавно вышедшей книге Джардайна и Сибсона [180] читатель найдет математическое обоснование методов, которыми пользуются в биологической таксономии; эти методы на самом деле имеют более широкое применение. Исследование этих методов продолжено в [175], [177], [178] и [179]. Ими же был предложен аксиоматический подход к кластерному анализу. Андерберг написал книгу по кластерному анализу, цель которой предложить унифицированный подход к кластерному анализу на элементарном уровне; в этой книге обсуждается также большое количество других вопросов кластерного анализа. Книга Сокала и Снита [336] служит хорошим справочным руководством, ориентированным на лиц, работающих в биологии; однако эта книга мало пригодна для исследователей других областей.

Болл [13] сделал прекрасный обзор и сравнение различных методов «поиска кластеров». Он разбивает все методы на семь групп: 1) вероятностные, 2) методы обнаружения сигнала, 3) кластеризации, 4) группировки, 5) собственных значений, 6) отыскания минимальной моды, 7) остальные. Наиболее употребительными методами, под которыми обычно и понимается кластерный анализ, являются методы кластеризации (3) и методы группировки (4). Ступенчатые методы, которые обсуждались в главе 1, — это методы группировки, а методы минимальной дисперсии той же главы принадлежат к группе методов кластеризации. Метод отыскания минимальной моды требует предварительного разбиения наблюдений на классы. Класс собственных значений Холла сродни факторному анализу и методу главных компонент многомерного анализа.

Другой класс, который включает в себя метод оценивания функции плотности, обсуждался в главе 5. Вероятностные методы Холла могут быть обобщены и на этот случай.

Методы кластеризации в общем случае наиболее эффективны; эти методы также хорошо поддаются интерпретации. Однако в исследованиях таксономии методы группировки более популярны. Исторически методы группировки были первыми методами кластеризации, которые впервые были применены в численной таксономии. Работа Сокала и Снита [336] может служить хорошим справочным руководством по этим методам.

Широкое признание нашел кластерный метод Болла и Холла [15], [16], [18] ИСОМАН (итеративный самообучающийся метод анализа данных). Этот метод был коротко упомянут в главе 1 как метод минимальной дисперсии. Он применялся, в частности, в задаче регистрации отдаленных объектов (мультиспектральные данные сканирования) в НАСА, в Центре пилотируе-

мых космических кораблей. Инструкцию по его эксплуатации можно найти в [166], [191], [193]. Окончательная версия ИСОМАНа была предложена Каном и Холли [194]. Применение ИСОМАН было показано на примере регистрации отдаленных объектов (мультиспектральные данные сканирования), который был описан в предыдущей главе.

Другим полезным методом является добавочный (adding) алгоритм Хартигена (записи лекций), который также описан у Кана и Холли [194]. Этот метод относится к классу разделительных (divise) иерархических алгоритмов. Кан и Холли обсуждают и другие методы, которые также применяются и весьма полезны (см. [325], [208], [404] и [302]).

Целью патерн-рекогносцировки, как и кластерного анализа, является разбиение данных на группы. Однако в патерн-анализе для каждого наблюдения известно, к какому классу информации оно принадлежит. Наджи [270] предлагает прекрасный обзор по патерн-рекогносцировке. Его статья насчитывает 148 ссылок.

Методы кластеризации могут быть разбиты на два больших класса: 1) разделительные, 2) агломеративные. Не надо только путать понятия алгоритма и метода. Данный метод может принадлежать либо к классу *разделительных,* либо к классу *агломеративных* методов. Разделительные методы разбивают множество объектов на группы, а агломеративные, наоборот, объединяют объекты в группы (кластеры). Разделительные методы были введены Эдвардсом и Кавалли-Сфорца [93], МакНотоном-Смитом и другими [235] и Ресчиньо и Макакоро [293]. Класс агломеративных методов шире; некоторые из этих методов были описаны в главе 1, [223], [234], [387]. Ступенчатый алгоритм кластеризации, рассмотренный в главе 1, является агломеративным алгоритмом. Агломеративные процедуры в общем случае не обязательно итеративные и предполагают существование правила объединения двух кластеров.

Некоторые методы, строго говоря, нельзя отнести ни к классу разделительных, ни к классу агломеративных; к таким методам относятся, например, методы, описанные в главе 1 [18], [33] и [237]. К этому же классу относится метод динамического программирования Дженсена [183], описанный в главе 4.

Очень мало сделано в области сравнения кластерных

методов. Очень трудно бывает определить, какой метод лучше. Выбор метода зависит как от целей исследования, так и от вида данных. Более того, в некоторых случаях вообще невозможно сравнивать два метода. Гауером [137] были сделаны сравнения методов Сокала и Миченера [334], Эвардса и Кавалли-Сфорца [93], Уильямса и Ламберта [394]. Рэнд [288] предлагает объективный критерий сравнения двух различных методов кластеризации. По другим вопросам применения кластерных методов см. Джардайн и Сибсон [180], глава 2, Борко [34] и Грин и Рао [144].

Фишер и Ван Несс [102] определили некоторые условия приемлемости, которые можно считать желательными почти во всех случаях кластерного анализа и сравнения этих свойств для некоторых стандартных методов. Условия приемлемости позволяют отбросить те процедуры, которые приводят к «плохому» разбиению на группы. Однако одновременно с исключением «плохих» методов применение этих условий в некоторых случаях приводит к отбрасыванию и пригодных методов кластеризации. Некоторые из условий приемлемости, рассматриваемые в [102], обсуждаются также Хартигеном [162], Джонсоном [186], Джардайном и Сибсоном [178]. Работа Фишера и Ван Несса [103], в которой рассматриваются проблемы приемлемости дискриминантного анализа, также представляет некоторый интерес.

Желаемое число кластеров, которые будут получены в результате применения того или иного метода кластеризации, может быть неизвестно. Часто это число не определяется из результатов кластеризации. В иерархических (агломеративных) и разделительных методах исследователь может найти нужное m из рассмотрения различных иерархических уровней. В динамическом подходе число m необходимо знать заранее. Методы паттерн-рекогносцировки также требуют предварительного знания значения m.

Следующий, достаточно неопределенный момент связан с понятием *сходимости* метода. Это понятие обсуждалось нами при рассмотрении некоторых методов *минимальной дисперсии* в главе 1. Результаты кластеризации должны быть единственными. Если данные мультимодальны, а моды определены достаточно хорошо, то результаты кластеризации имеют тенденцию быть един-

ственными. Если данные представляют собой *плоскую поверхность*, т. е. такие, какие получаются, если выборку производить из равномерного распределения, то результаты кластеризации не обязательно будут единственными. Это приведет к нечеткому разбиению на кластеры, и практически результатом разбиения будет один кластер, содержащий все объекты ($m=1$).

Методы, которые мы обсуждали в этой монографии, предназначались для кластеризации объектов или элементов. Однако эти же методы можно применять и для кластеризации признаков или характеристик. В [161] Хартиген предлагает метод двойной кластеризации, т. е. метод, который кластеризует и по объектам и по признакам одновременно.

Кластерный анализ тесно связан с другими методами многомерного анализа, методом главных компонент, дискриминантным анализом, факторным анализом.

Дискриминантный анализ предназначен для получения предварительной классификации данных. Перегруппировывая данные и вычисляя новое значение дискриминантной функции в результате итераций, приходят к «наилучшему» разбиению данных на группы. Касетти [45], Ханг и Дюбс [170] описывают программу применения этого метода. Мейер [261] предлагает аналогичный метод, в котором пользуются одной характеристикой, представляющей наибольший интерес. В более поздней статье Мейер [263] рассматривает метод, при котором матрица наблюдений предварительно сокращается ($p'<p$) с помощью метода главных компонент, после чего для построения кластеров вводится критерий расстояния и линейная дискриминационная функция. Урбах [375] предложил метод дискриминантного анализа разбиения разнородной многомерной совокупности на группы; разбиения повторяются до тех пор, пока не будет получен удовлетворительный результат, который записывается в терминах вероятности «плохой» классификации.

ЛИТЕРАТУРА

[1] Abraham C. T. A note on a measure of similarity used in the DICO experiment, Appendix I, Quarterly Report 3, Vol. I, Contract AF 19(626)—10.

[2] Abraham C. T. Evaluation of clusters on the basis of random graph theory, Yorktown Heights, N. Y.: IBM Corporation, IBM Res. Memo. Nov., 1962.

[3] Adhikari B. P. and Joshi D. D. Distance discrimination et resume exhaustif, Pbls. Inst. Statist., Vol. 5, (1956), 57—74.

[4] Aitken M. A. The correlation between variate values and ranks in a doubly truncated normal distribution, Biometrika, Vol. 53, Parts 1/2, (1966), 281—282.

[5] Anderberg M. R. Cluster analysis for applications, (in press), Dec., 1971.

[6] Anderberg M. R. An Annotated Bibliography of Cluster Analysis, Mechanical Engineering Department, University of Texas at Austin (in preparation), (1972).

[7] Anderson T. W. An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., New York, (1958).

Андерсон Т. Введение в многомерный статистический анализ. М., Физматгиз, 1963.

[8] Archer W. B. Computation of Group Job Descriptions from Occupational Survey Data, Report Number PRL-TR-66-12, Personnel Research Laboratory, Lackland AFB, Texas, 31 pp.

[9] Armstrong J. S. and Soelberg P. On the interpretation of factor analysis, Psychol. Bull., 70 (1968), 361—364.

[10] Astrahan M. M. Speech Analysis by Clustering, or the Hyperphoneme Method, Stanford Artificial Intelligence Project Memo AIM-124, Stanford University, 22 pp., (1970).

[11] Balas E. An additive algorithm for solving linear programs with zero-one variables, Operations Res., 13 (1965), 517—546.

[12] Balinski M. L. Integer programming: methods, uses and computation, Management Sci., 12 (Nov. 1965), 253—313.

[13] Ball G. H. A Comparison of Some Cluster — Seeking Techniques, Report Number RADC-TR-66-514, Stanford Research Inst. Menlo Park, California, 47 pp., (1966).

[14] Ball G. H. Classification Analysis, Technical Note, Stanford Research Inst. Menlo Park, California, (1970).

[15] Ball G. H. Data Analysis in the social sciences — what about details? American Federation of Information Processing Societies Conference Proceedings: 1965 Fall Joint Computer Conference, 27 (1965), Part 1, 533—560, (Washington: Spartan Books; London: Macmillan).

[16] Ball G. H. and Hall D. J. A clustering technique for summarizing multivariate data, Behavioral Sciences, Vol. 12, No. 2, (Mar., 1967), 153—155.
[17] Ball G. H. and Hall D. J. Background information on clustering techniques, Stanford Research Inst., (Jul., 1968).
[18] Ball G. H. and Hall D. J. ISODATA, A. Novel Method of Data Analysis and Pattern Classification, Technical Report, Menlo Park, California: Stanford Research Inst., 72 pp., (1965).
[19] Ball G. H. and Hall D. J. PROMENADE — An On-Line Pattern Recognition System, Report Number RADC-TR-67-310, Stanford Research Inst., 124 pp., (1967).
[20] Baker F. B. Latent class analysis as an association model for information retrieval, in Stevens, Giuliano and Heilprin (eds.), Statistical Association Methods for Mechanized Documentation, National Bureau of Standards Miscellaneous Publication Number 269, U. S. Government Printing Office, Washington, D. C., (1965), 149—155.
[21] Bartels P. H., Bahr G. F., Calhoun D. W. and Wied G. L. Cell recognition by neighborhood grouping technique in TICAS, Acta Cytologica. Vol. 14, No. 6, (1970), 313—324.
[22] Barton D. E. and David F. N. Spearman's 'Rho' and the matching problem, Brit. J. Statist. Psychol., 9 (1956), 69—73.
[23] Bass B. M. Iterative inverse factor analysis — a rapid method for clustering persons, Psychometrika, Vol. 22, No. 1, (Mar., 1957), 105—107.
[24] Baxendale P. An empirical model for computer indexing, Machine Indexings Progress and Problems, American University, Washington, D. C., (Febr. 13—17, 1961), 267.
[25] Beale E. M. L. Euclidean cluster analysis, Bull. I. S. I., 43, 2 (1969), 92—94.
[26] Beale E. M. L. Selecting an optimum subset, in J. Abadie (ed.), Integer and Nonlinear Programming, Amsterdam: North Holland Publishing Company, (1970).
[27] Bellman R. E. and Dreyfus S. E. Applied Dynamic Programming, Princeton, N. J.: Princeton University Press, (1962).
Беллман Р., Дрейфус С. Прикладные задачи динамического программирования. М., «Наука», 1965.
[28] Benders J. F. Partitioning procedures for solving mixed-variables programming problems, Numerische Mathematik, 4, (Febr. 1962), 238—252.
[29] Bhattacharyya A. On a measure of divergence between two statistical populations defined by their probability distributions, Bull. Calcutta Math. Soc., Vol. 35, (1943), 99—109.
[30] Birnbaum A. and Maxwell A. E. Classification procedures based on Bayes' formula, in L. J. Cronbach and Goldine C. Gleser (eds.), Psychological tests and personnel decisions, Urbana: University of Illinois Press, (1965).
[31] Bledsoe W. W. A. corridor — projection method for determining orthogonal hyperplanes for pattern recognition, unpublished report, Panoramic Research Cor., Palo Alto, California (1963).

[32] Block H. D., Knight B. W. and Rosenblatt F. The perception: A model for brain functioning, II, Rev. Modern Phys., Vol. 34, No. 1, (Jan. 1962), 135—142.
[33] Bonner R. E. On some clustering techniques, IBM Journal, 22, (Jan. 1964), 22—32.
[34] Borko H., Blankenship D. A. and Burket R. C. On — Line Information Retrieval Using Associative Indexing, RADC-TR-68-100, Systems Development Corporation, (1968), 124 pp.
[35] Bottenberg R. A. and Christal R. E. An Iterative Technique for Clustering Criteria Which Retains Optimum Predictive Efficiency. WADD-TN-61-30, Lackland AFB, Texas: Personnal Research Laboratory, Wright Air Development Division, (Mar., 1961).
[36] Boulton D. M. and Wallace C. S. A program for numerical classification, Comput. J., Vol. 13, (1970), 63—69.
[37] Bradley R. A., Katti S. K. and Coons I. J. Optimal scaling for ordered categories, Psychometrika, Vol, 27, No. 4, (1962), 355—374.
[38] Brennan E. J. An analysis of the adaptive filter, General Electric Elec. Lab. Tech. Information Series Report R61 ELS-20, Syracuse, N. Y., (1961).
[39] Bryan J. G. Calibration of Qualitative or Quantitative Variables for Use in Multiple — Group Discriminant Analysis, The Travelers Weather Research Center, Hartford, Conn., 26 pp.
[40] Bryan J. K. Classification and clustering using density estimation, Ph. D. Dissertation, University of Missouri, Columbia, Missouri, (Aug., 1971).
[41] Butler G. A. A vector field approach to cluster analysis, Pattern Recognition, 1 (1969), 291—299.
[42] Cacoullos T. Estimation of multivariate density, Ann. Inst. Statist. Math., Vol. 18, (1966), 179—189.
[43] Campbell J. P. A hierarchical cluster analysis of the core courses in an engineering curriculum, J. Exp. Educ., (1966), 35, 63—69.
[44] Carroll J. B. The nature of the data, or how to choose a correlation coefficient, Psychometrika, Vol. 26, No. 4, (1961), 347—372.
[45] Casetti E. Classificatory and Regional Analysis by Discriminant Iterations, TR-12, Contract Nonr — 1228 (26), Northwestern University, Evanston, III., 99 pp., (1964).
[46] Castellan N. J. Jr. On the estimation of the tetrachoric correlation coefficient, Psychometrika, Vol. 31, No. 1, (1966), 67—73.
[47] Cattell R. B. A note on correlation clusters and cluster search methods, Psychometrika, (Sept., 1944), 9, 169—184.
[48] Cattell R. B. Factor analysis: An introduction to essentials II. The role of factor analysis in research, Biometrics, Vol. 21, No. 2, (1965), 405—435.
[49] Cattell R. B. and Coulter M. A. Principles of behavioural taxonomy and the mathematical basis of the taxonomic computer program, Brit. J. Math. Statist, Psychol., 19 (1966), 237—269.

[50] Charnes A. A. and Cooper W. W. Management Models and Industrial Applications of Linear Programming. Vol. 1, New York: John Wiley & Sons, Inc., (1961).
[51] Chernoff H. Metric Considerations in Cluster Analysis, Technical Report No. 67, Department of Statistics, Stanford University, Stanford, California, 16 pp., (1970).
[52] Christal R. E. and Ward J. H. Jr. Applications of a new clustering technique which minimizes loss in terms of any criterion specified by the investigator, paper presented at the meeting of the American Psychological Association, New York City, (Sept., 1961).
[53] Christal R. E. and Ward J. H. Jr. The MAXOF clustering model. Lackland AFB, Texas: Personnel Research Division Air Force Human Resources Laboratory (AFSC), (1970) in press.
[54] Christal R. E. and Ward J. H. Jr. Use of an objective function in clustering people or things into mutually exclusive categories, paper presented at Conference on Cluster Analysis of Multivariate Data, New Orleans, La., (Dec., 1966).
[55] Clark P. J. An extension of the coefficient of divergence for use with multiple characters, Copeia 2 (1952), 61—64.
[56] Cochran W. G. and Hopkins C. E. Some classification problems with multivariate qualitative data, Biometrics, Vol. 17, No. 1, (1961), 10—32.
[57] Cole A. J. Numerical Taxonomy, Academic Press, New York, (1969).
[58] Cole L. C. The measurement of interspecific association, Ecology 30 (1949), 411—424.
[59] Cole L. C. The measurement of partial interspecific association, Ecology 38 (1957), 226—233.
[60] Constantinescu P. A method of cluster analysis, Brit. J. Math. Statist. Psychol. 20(1), (1967), 93—106.
[61] Cooper D. B. Nonsupervised adaptive signal detection and pattern recognition, Raytheon Report, (Oct. 22, 1963).
[62] Cooper D. B. and Cooper P. W. Adaptive pattern recognition and signal detection without supervision, IEEE International Convention Record. Part 1, (1964), 246—256.
[63] Cooper D. B. and Cooper P. W. Nonsupervised adaptive signal detection and pattern recognition, Information and Control, Vol. 7, No. 3, (Sept., 1964).
[64] Cooper W. W. and Majone G. A description and some suggested extensions for methods of cluster analysis, Internal Working Memorandum, Carnegie — Mellon University.
[65] Cover T. M. and Hart P. E. Nearest-neighbor pattern classification, IEEE Trans. Inf. Theory, 13, (1967), 21—27.
[66] Cox D. R. Note on grouping, J. Amer. Statist. Assoc., 52, (1957), 543—547.
[67] Cramer H. On the composition of elementary errors, Skand. Aktuarietids, Vol. 11, (1928), 13—74 and 141—180.
[68] Cramer H. The Elements of Probability Theory and Some of its Applications, John Wiley & Sons, Inc., New York, (1946).
[69] Crawford R. M. M. and Wishart D. A rapid classification and ordination method and its application to vegetation mapping, J. Ecology, Vol. 56, No. 2, (1968), 385—404.

[70] Crawford R. M. M. and Wishart D. A rapid multivariate method for the detection and classification of groups of ecologically related species, J. Ecology, Vol. 55, No. 2, (1967), 505—524.
[71] Cronbach L. J. and Gleser G. C. Assessing the similarity between profiles, Psychol. Bull., Vol. 50, No. 6, (1953), 456—473.
[72] Dagnelie P. On different methods of numerical classification, Rev. Statist. Appl. 14, III (1966), 55—75.
[73] Daly R. F. Adaptive binary detection, Stanford Elec. Lab. Tech. Report No. 2003—2, Stanford, California, (Jun. 26, 1961).
[74] Daly R. F. The adaptive binary-detection problem on the real line, Stanford Elec. Lab. Report SEL-62-030, Stanford, California, (Febr., 1962).
[75] Daniels H. H. Rank correlation and population models, J. Roy. Statist. Soc., B 12 (1950), 171—181.
[76] Darling D. A. The Kolmogorov-Smirnov, Cramer-Von Mises tests, Ann. Math. Statist., Vol. 28, (Dec. 1957), 823—838.
[77] Day N. E. Estimating the components of a mixture of normal distributions, Biometrika, Vol. 56, (1969), 463—474.
[78] David F. N., Barton D. E., Ganeshalingham S., Harter H. L., Kim P. J., Merrington M. and Walley D. Normal Centroids, Medians and Scores for Original Data, Cambridge University Press, Cambridge, England, (1968).
[79] David S. T., Kendall M. G. and Stuart A. Some questions of distribution in the theory of rank correlation, Biometrika 38 (1951), 131—140.
[80] Demiremen F. Multivariate Procedures and FORTRAN IV Program for Evaluation and Improvement of Classifications, Computer Contribution 31, State Geological Survey, The University of Kansas, Lawrence, 51 pp., (1969).
[81] Dempster A. P. Elements of Continuous Multivariate Analysis, Reading, Massachusetts: Addison-Wesley Publishing Co., (1969).
[82] Dice L. R. Measures of the amount of ecological association between species, Ecology 26 (1945), 297—302.
[83] Dobes J. Partitioning algorithms, Inf. Processing Math., 13, (1967), 307—313.
[84] Doyle L. B. Breaking the Cost Barrier in Automatic Classification, Professional Paper SP—2516, System Development Corp., Santa Monica, California, 62 pp., (1966).
[85] Dubes R. C. Information Compression, Structure Analysis and Decision Making with a Correlation Matrix, Michigan State University, East Lansing, Michigan, 243 pp., (1970).
[86] Dubin R. Typology of Empirical Attributes: Multi-dimensional Typology Analysis (MTA), TR-5, University of California at Irvine, 17 pp., (1971).
[87] Dubin R. and Champoux J. E. Typology of Empirical Attributes: Dissimilarity Linkage Analysis (DLA), Technical Report 3, University of California at Irvine, 31 pp., (1970).

[88] Duncan D. B. Multiple range and multiple F tests, Biometrics, Vol. 11, No. 1, (1955), 1—42.
[89] Durbin J. and Stuart A. Inversions and rank correlation coefficients, J. Roy. Statist. Soc., B, 13 (1951), 303—309.
[90] Eades D. C. The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance, Systematic Zoology, Vol. 14, No. 2, (1965), 98—100.
[91] Eddy R. P. Class Membership Criteria and Pattern Recognition, Report 2524, Naval Ship Research and Development Center, Washington, D. C., 47 pp., (1968).
[92] Edwards A. W. F. The measure of association in a 2×2 table, J. Roy Statist. Soc., Series A, Vol. 126, Part I, (1963), 109—114.
[93] Edwards A. W. F. and Cavalli—Sforza L. L. A method for cluster analysis, Biometrics, Vol. 21, No. 2, (1965), 362—375.
[94] Eisen M. Elementary Combinatorial Analysis, Gordon and Breach Science Publishers, (1969).
[95] Elkins T. A. Cubical and spherical estimation of multivariate probability density, J. Amer. Statist. Assoc., Vol. 63, No. 324, (1968), 1495—1513.
[96] Engleman L. and Hartigan J. A. Percentage points of a test for clusters, J. Amer. Statist. Assoc., Vol. 64, (1969), 1947—1948.
[97] Ericson W. A. A note an partitioning for maximum between sum of squares, Appendix C in J. A. Sonquist and J. N. Morgan, The Detection of Interaction Effects, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.
[98] Estabrook G. F. A mathematical model in graph theory for biological classification, J. Theoret. Biol., 12 (1966), 297—310.
[99] Eusebio J. W. and Ball G. H. ISODATA—LINES—A program for describing multivariate data by piecewise—linear curves, Proceedings of International Conference on Systems Science and Cybernetics, University of Hawaii, Honolulu, Hawaii, (Jan. 1968), 560—563.
[100] Farris J. S. On the Cophenetic Correlation coefficients, Systematic Zoology, Vol. 18, No. 3, (1969), 279—285.
[101] Firschein O., Fischler M. Automatic subclass determination for pattern recognition applications, Trans. PGEC, EC-12, No. 2 (Apr., 1963).
[102] Fisher L. and Van Ness J. Admissible clustering procedures, Biometrika 58, (1971), 91—104.
[103] Fisher L. and Van Ness J. Admissible discriminant analysis, J. Amer. Statist. Assoc. 68 (1973), 603—607.
[104] Fisher R. A. The precision of discriminant functions, Ann. Eugenics, Vol. 10, (1940), 422—429.
[105] Fisher R. A. Statistical Methods for Research Workers, Hafner Publishing Co., New York, Thirteenth Edition, reprint, (1963).
Фишер Р. А. Статистические методы для исследователей. М., Госстатиздат, 1958.
[106] Fisher R. A. The use of multiple measurements in taxonomic problems, Ann. Eugenics, Vol. VII, Part II, (1936), 179—188.
[107] Fisher R. A. and Yates F. Statistical Tables for Biological, Agricultural and Medical Research, Oliver and Boyd, London, (1953).
[108] Fisher W. D. Clustering and Aggregation in Economics, The Johns Hopkins Press, Baltimore, Maryland, (1968).
[109] Fisher W. D. On a pooling problem from the statistical decision viewpoint, Econometrika, (1953), 21, 567—585.
[110] Fisher W. D. On grouping for maximum homogeneity, J. Amer. Statist. Assoc., Vol. 53, (1958), 789—798.
[111] Fisher W. D. Simplification of economic models, Econometrika, Vol. 34, No. 3, (Jul., 1966), 563—584.
[112] Flake R. H. and Turner B. L. Numerical classification for taxonomic problems, J. Theoret. Biol., Vol. 20 (1968), 260—270.
[113] Forgy E. W. Classification so as to relate to outside variables, in M. Lorr and S. B. Lyerly (eds.), Final Report, Conference on Cluster Analysis of Multivariate Data, Catholic University of America, Washington, D. C., (1966), pp. 13.01—13.12.
[114] Forgy E. W. Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications, paper presented at Biometric Society meetings. Riverside, California, (abstract in Biometrics. Vol. 21, No. 3), (1965), p. 768.
[115] Forgy E. W. Detecting 'Natural' Clusters of individuals, Western Psychological Association Meetings, Santa Monica. California, (Apr. 19, 1963).
[116] Forgy Edward W. Detecting 'Natural' clusters of individuals, report at Western Psychological Association, Dept, of Psychiatry, University of California Medical Center, Los Angeles, California, (Apr., 1963), 1—10.
[117] Forgy E. W. Evaluation of several methods for detecting sample mixtures from different N-dimensional populations, American Psychology Association Meetings, Los Angeles, Calif., (Sept. 9, 1964). (Available from author at Center for Health Sciences, U. C. L. A., Los Angeles, Calif.).
[118] Fortier J. J. Contributions to item selection, Tech. Rep. no. 2, Laboratory for Quantitative Research in Education, Stanford University, (1962).
[119] Fortier J. and Solomon H. Clustering procedures, Multivariate Analysis, ed. by P. R. Krishnaiah, Academic Press, N. Y., (1966), 493—506.
[120] Fralick S. C. The synthesis of machines which learn without a teacher, IEEE Trans. on Information Theory, Vol. IT-13, (1967), 57—64.
[121] Fralick S. C. The synthesis of machines which learn without a teacher, Tech. Report No. 6103-8, Stanford University, (Apr., 1964).
[122] Friedman H. P. and Rubin J. On some invariant criteria for grouping data, J. Amer. Statist. Assoc., Vol. 62, (1967), 1159—1178.
[123] Froemel E. C. A comparison of computer routines for the calculation of the tetrachoric correlation coefficient, Psychometrika, Vol. 36, No. 2, (1971), 165—174.

[124] Fu K. S., Langrebe D. A. and Phillips T. L. Information processing of remotely sensed agricultural data, Proc. IEEE, Vol. 57, No. 4, (Apr., 1969), 639—653.
[125] Fukunaga K. and Koontz W. L. A criterion and an algorithm for grouping data, IEEE Trans. On Computers, Vol. C-19, (Oct. 1970), 917—923.
[126] Garfinkel R. S. and Nemhauser G. L. Optimal political districting by implicit enumeration techniques, Management Sci., Vol. 16, No. 8, (1970), B495—B508.
[127] Garfinkel R. and Nemhauser G. L. The set-partitioning problem: Set covering with equality constraints, Operations Res., 17, (Sept. — Oct. 1969), 848—856.
[128] Gengerelli J. A. A method for detecting subgroups in a population and specifying their membership, J. Psychology, Vol. 55, (1963), 457—468.
[129] Gitman I. and Levine M. D. An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique, IEEE Trans. on Comput., Vol. C-19, No. 7, (1970), 583—593.
[130] Glaser E. M. Signal detection by adaptive filters, IRE Trans. on Info. Theory, Vol. IT-7, No. 2, (Apr., 1961).
[131] Gleason A. M. A search problem in the n-cube, Proc. Symposium in Appl. Math. Amer. Math. Soc., 10, (1960). 175—178.
[132] Glover F. A multiphase-dual algorithm for the zero—one integer programming algorithm. Operations Res., Vol. 13. No. 6, (Nov. — Dec. 1965), 879—919.
[133] Goldberger A. S. Impact Multipliers and Dynamic Properties of the Klien—Goldberger Model, Amsterdam: North—Holland Publishing Co., (1959).
[134] Gomory R. E. All-integer integer programming algorithm, in J. L. Muth and G. L. Thompson (eds.), Industrial Scheduling, Englewood Cliffs, N. J.: Prentice—Hall (1963).
[135] Goodman L. A. and Kruskal W. H. Measures of association for cross classifications, J. Amer. Statist. Assoc., Vol. 49, (1954), 732—764.
[136] Goodman L. A. and Kruskal W. H. Measure of association for cross-classification, II, J. Amer. Statist. Assoc., Vol. 54, (1959), 123—163.
[137] Gower J. C. A comparison of some methods of cluster analysis, Biometrika, Vol. 23, No. 4, (1967), 623—637.
[138] Gower J. C. Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika, Vol. 53, No. 3/4, (1966), 325—338.
[139] Gower J. C. and Ross G. J. S. Minimum spanning trees and single linkage cluster analysis, Appl. Statist., Vol. 18, No. 1, (1969), 54—64.
[140] Grason J. Methods for the Computer—Implemented Solution of a Class of «Floor—Plan» Design Problems, Ph. D. Dissertation, Electrical Engineering Department, Carnegie—Mellon University, Pittsburgh, Pennsylvania, 374 pp. (1970).
[141] Gray H. L. and Schucany W. R. The Generalized Jackknife Statistic, New York: Marcel Dekker, Inc., (1972).
[142] Green P. E. and Carmone F. J. Multi-dimensional Scaling and Related Techniques in Marketing Analysis, Allyn and Bacon Inc., Boston, (1970).

[143] Green P. E., Frank R. E. and Robinson P. J. Cluster analysis in test market selection, Management Sci., 13, (1967), 13—387—400.
[144] Green P. E. and Rao V. R. A note on proximity measures and cluster analysis, J. Marketing Research, Vol. VI, (1969), 359—364.
[145] Hadley G. Nonlinear and dynamic programming, Addison—Wesley, Reading, Massachusetts, (1964).
[146] Haggard E. A. Intra-Class Correlation and the Analysis of Variance, Dryden, N. Y., (1958).
[147] Hall A. V. Avoiding informational distortions in automatic grouping programs, Systematic Zoology, Vol. 18, No. 3, (1969), 318—329.
[148] Hall D. J., Ball G. H., Wolf D. E. and Eusebio J. PROMENADE: An Improved Interactive-Graphics Man/Machine System for Pattern Recognition, Report Number RADC-TR-68-572, Stanford Research Institute, Menlo Park, Calif. 173 pp., (1969).
[149] Hamdan M. A. Estimation of class boundaries in fitting a normal distribution to a qualitative multinomial distribution, Biometrics, Vol. 27, No. 2, (1971), 457—459.
[150] Hamdan M. A. On the polychoric method for estimation of [Rho] in contingency tables, Psychometrika, Vol. 36, No. 3, (1971), 253—259.
[151] Hammer P. L. and Rudeanu S. Boolean Methods in Operations Research and Related Areas, New York: Springer-Verlag, (1968).
[152] Hanson N. R. Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science, Cambridge University Press, New York, (1958).
[153] Haralick R. M. and Dinstein I. An iterative clustering procedure, IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-1, No. 3, (Jul., 1971), 275—289.
[154] Harary F. Graph Theory, Addison-Wesley, Reading, Massachusetts, (1969).
[155] Harding E. F., The number of partitions of a set of N points in k dimensions induced by hyperplanes, Proc. Edinburg Math. Soc., Vol. 15, (1967), 285—289.
[156] Harding E. F. The probabilities of rooted tree—shapes generated by random bifurcation, Advances in Appl. Probability, Vol. 3, No. 1, (1971), 44—77.
[157] Harman H. H. Modern Factor Analysis, University of Chicago Press, Chicago, Illinois, (1960).
Харман Г. Современный факторный анализ. М., «Статистика», 1972.
[158] Harrison J. Cluster Analysis, Metra, 7, (1968), 513—518.
[159] Harter H. L. Expected values of normal order statistics, Biometrika, Vol. 48, Part 1/2, (1961), 151—165.
[160] Hartigan J. A. Clustering a Data Matrix, working paper, Department of Statistics, Yale University, New Haven, Connecticut, (1970), 55 pp.
[161] Hartigan J. A. Direct clustering of a data matrix, J. Amer. Statist. Assoc., Vol. 67, (1972), 123—129.
[162] Hartigan J. A. Representation of similarity matrices by trees, J. Amer. Statist. Assoc., Vol. 62, (1967), 1140—1158.

[163] Hartigan J. A. Using subsample values as typical values, J. Amer. Statist. Assoc., Vol. 64, (1969), 1303—1317.
[164] Hasselblad V. Estimation of parameters for a mixture of normal distributions, Technometrics, Vol. 8, (1966), 431—446.
[165] Henschke C. I. Manpower Systems and Classification Theory, Ph. D. Dissertation, University of Georgia, Athens, Georgia, 261 p., (1969), cited in Dissertation Abstracts, Vol. 30, No. 8, p. 3911-B.
[166] Holley W. A. Description and user's guide for the IBM 360/44 ISODATA PROGRAM, Lockheed Electronics Co., Inc., HASD, Houston, Texas, Tech. Rep. 640-TR-030, (Sept., 1971).
[167] Holmes R. A. and MacDonald R. B. «The physical basis of systems design for remote sensing in agriculture», Proceedings IEEE, Vol. 57, Apr. 1969, p. 629—639.
[168] Holzinger K. J. and Harman H. H. Factor Analysis, Chicago Press, Chicago, Ill., (1941).
[169] Huang F. Per field classifier for agriculture applications, LARS Information Note 060569, Purdue University, Lafayette, Indiana, (Jun., 1969).
[170] Hung A. Y. and Dubes R. C. An Introduction to Multiclass Pattern Recognition in Unstructured Situations, Interim Scientific Report No. 12, Division of Engineering Research, Michigan State University, East Lansing, Michigan, 66 pp., (1970).
[171] Hyvarinen L. Classification of Qualitative Data, Brit. Info. Theory J., (1962), 83—89.
[172] Isaacson E. and Keller H. B. Analysis of Numerical Methods, John Wiley and Sons, Inc., New York, (1966).
[173] Jaccard P. Nouvelles Recherches sur la distribution florale, Bull. Soc. Vand. Sci. Nat. 44, (1908), 223—270.
[174] Jancey R. C. Multidimensional group analysis, Australian J. Botany, Vol. 14, No. 1, (1966), 127—130.
[175] Jardine C. J., Jardine N. and Sibson C. The structure and construction of taxonomic hierarchies, Mathematical Biosciences, Vol. 1, No. 2, (1967), 173—179.
[176] Jardine N. Algorithm, methods, and models, in the simplification of complex data, Comput. J., 13, (1970), 116—117.
[177] Jardine N. Towards a general theory of clustering, Biometrics, 25, (1969), 609—610.
[178] Jardine N. and Sibson R. The construction of hierarchic and nonhierarchic classifications, Comput. J., Vol. 11, (1968), 177—184.
[179] Jardine N. and Sibson R. A model for taxonomy, Math. Biosci., 2, (1968), 465—482.
[180] Jardine N. and Sibson R. Mathematical Taxonomy, John Wiley and Sons, New York, (1971).
[181] Jeffreys H. Theory of probability, Oxford University Press, (1948).
[182] Jeffreys H. An invariant for the prior probability in estimation problems, Proc. Roy. Soc, A., Vol. 186, (1946), 454—461.
[183] Jensen R. E. A dynamic programming algorithm for cluster analysis, Operations Res., 12, (Nov.—Dec. 1969), 1034—1057.
[184] John S. On identifying the population of origin of each observation in a mixture of observations from two normal populations, Technometrics, Vol. 12, (1970), 553—565.

[185] Johnson P. O. The quantification of qualitative data in discriminant analysis, J. Amer. Statist. Assoc., Vol. 45, (1950), 65—76.
[186] Johnson S. C. Hierarchical clustering schemes, Psychometrika, Vol. 32, No. 3, (Sept. 1967), 241—254.
[187] Kahl J. A. and Davis J. A. A comparison fo indexes of socio-economic status, Amer. Sociol. Rev., Vol. XXI, 3 (1956).
[188] Kailath T. The divergence and Bhattacharyya distance measures in signal selection, IEEE Trans. on Comm. Tech., Vol. COM-15, (Febr. 1967), 52—60.
[189] Kaminuma T., Takekawa T. and Watanabe S. Reduction of clustering problem to pattern recognition, Pattern Recognition, Vol. 1, (1969), 195—205.
[190] Kan E. P. F. Data clustering: An overview, Lockheed Electronics Co., Inc., HASD, Houston, Texas, Tech. Rep. 640-TR-080, (Mar., 1972).
[191] Kan E. P. F. ISODATA: Thresholds for splitting clusters, Lockheed Electronics Co., Inc., HASD, Houston, Texas, Tech. Rep. 640-TR-058, (Jan., 1972).
[192] Kan E. P. F. On an iterative clustering technique, Lockheed Electronics Co., Inc., HASD, Houston, Texas, Tech. Memo. LEC TM642-214, (Nov., 1971).
[193] Kan E. P. F. and Holley W. A. Experience with ISODATA, Lockheed Electronics Co., Inc., HASD, Houston, Texas, Tech. Memo TM 642-354, (Mar., 1972).
[194] Kan E. P. F. and Holley W. A. More on clustering techniques with final recommendations on ISODATA, Lockheed Electronics Co., Inc., HASD, Houston, Texas, Tech. Rep. LEC 640-TR-112, (May, 1972).
[195] Kaskey G. et al. Cluster formation and diagnostic significance in psychiatric symptom evaluation, Proc. Fall Jt. Computer Conf., (1962), p. 285.
[196] Kazmierezak H. and Steinbuch K. Adaptive systems in pattern recognition, IEEE Trans. on Electronic Computers, Vol. EC-12, No. 6, (Dec., 1963).
[197] Keifer J. and Wolfowitz J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Ann. Math. Statist., Vol. 27, (1956), 887—906.
[198] Kendall M. G. A Course in Multivariate Analysis, Hafner Publishing Company, New York, Fourth Impression, (1968).
[199] Kendall M. G. A new measure of rank correlation, Biometrika, 30 (1938), 81—93.
[200] Kendall M. G. Discrimination and classification, Multivariate Analysis, ed. by P. R. Krishnaiah, Academic Press, N. Y., (1966), 165—184.
[201] Kendall M. G. Rank Correlation Methods, Griffin, London, 2nd edition, (1965).
Кендэл М. Ранговые корреляции. М., «Статистика», 1975.
[202] Kendall M. G., Kendall S. F. H. and Smith B. B. The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times, Biometrika 30 (1938), 251—273.
[203] Kendall M. G. and Stuart A. The Advanced Theory of Statistics, Vol. II, Inference and Relationship, Charles Griffin & Co., Ltd., London, (1961).

Кендалл М. Дж., Стьюарт А. Статистические выводы и связи. М., «Наука», 1973.
[204] King B. Stepwise clustering procedures, J. Amer. Statist. Assoc., (1967), 62, 86—101.
[205] Kochen M. Techniques for information retrieval research: State of the art, presented at IBM World Trade Corporation Information Retrieval Symposium at Blaricum, Holland, (Nov., 1962), to be published in the proceedings of the symposium.
[206] Kochen M. and Wong E. Concerning the possibility of a cooperative information exchange, IBM Journal of Research and Development, Vol. 6, No. 2, (Apr., 1962), 270—271.
[207] Kolmogorov A. N. Sulla determinazionne empirica di une legge di distribuzione, Giorn, dell'Instit. degli att., Vol. 4, (1933), 83—91.
[208] Koontz W. L. and Fukunaga K. A nonparametric valley-seeking technique for clustering analysis, IEEE Trans. on Computers, Vol. C-21, No. 2, (Febr., 1972), 171—178.
[209] Korn G. A. and Korn T. M. Mathematical Handbook for Scientists and Engineers, New York: McGraw-Hill Book Company, (1968).
Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. М., «Наука», 1968.
[210] Kraft C. H. Some conditions for consistency and uniform consistency of statistical procedures, University of California Publications in Statistics, (1955).
[211] Kruskal J. B. Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis, Psychometrika, 29 (1964), 1—27.
[212] Kruskal J. B. Nonmetric multidimensional scaling: A numerical method, Psychometrika, 29, No. 2, (Jun., 1964), 115—129.
[213] Kruskal J. B. Jr. On the shortest spanning subtree of a graph and the traveling salesman problem, Proc. Amer. Math. Soc., No. 7, (1956), 48—50.
[214] Kruskal W. H. Ordinal measures of association, J. Amer. Statist. Assoc., Vol. 53, (1958), 814—861.
[215] Kshirsagar A. M. Goodness of fit of an assigned set of scores for the analysis of association in a contingency table, Ann. Inst. Statist. Math., Vol. 22, No. 2, (1970), 295—306.
[216] Kullback S. and Liebler R. A. On information and sufficiency, Ann. Math. Statist., Vol. 22, (1951), 79—86.
[217] Kullback S. Information Theory and Statistics, New York, Dover Publications, Inc., (1968).
Кульбак С. Теория информации и статистика. М., «Наука», 1967.
[218] Labovitz S. In defense of assigning numbers to ranks, American Sociological Review, Vol. 36, No. 3, (1971), 520—521.
[219] Labovitz S. Some observations on measurement and statistics, Social Forces, Vol. 46, No. 2, (1967), 151—160.
[220] Lancaster H. O and Hamdan M. A. Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters, Psychometrika, Vol. 29, No. 4, (1964). 383—391.
[221] Lance G. N. and Williams W. T. Computer programs for hierarchical polythetic classification ('Similarity Analyses'), Comput. J., Vol. 9, No. 1, (1966), 60—64.
[222] Lance G. N. and Williams. W. T. Computer program for monothetic classification ('Association Analysis'), Comput. J., Vol. 8, No. 3, (1965), 246—249.
[223] Lance G. N. and Williams W. T. A general theory of classificatory sorting strategies. I. Hierarchical systems, Comput. J., Vol. 9, No. 4, (1967), 373—380.
[224] Lance G. N. and Williams W. T. A general theory of classificatory sorting strategies II. Clustering systems, Comput. J., Vol. 10, No. 3, (1967), 271—276.
[225] Lance G. N. and Williams W. T. A generalized sorting strategy for computer classifications, Nature, Vol. 212, (1966), p. 218.
[226] Landgrebe D. A. and LARS Staff LARSYAA, A Processing system for airborne earth resource data, LARS Information Note 091968, Purdue University, Lafayette, Indiana, Sept. 1969.
[227] Landgrebe D. A. and Phillips T. L. A multichannel image data handling system for agriculture remote sensing, Proc. Seminar on Computerized Image Handling Techniques, Washington, D. C., Jun., 1967, pp. XIT-1 to 10.
[228] Lemke C. E. and Spielberg K. Direct search algorithms for zero-one and mixed integer programming, Operations Res., Vol. 15, No. 5, (Sept. — Oct., 1967), 892—914.
[229] Levine M. D. Feature Selection: a survey, Proc. IEEE, Vol. 57, No. 8, Aug. 1969, 1391—1408.
[230] Lewis P. M. The characteristic selection problem in recognition systems, IRE Transaction on Information Theory, IT-8, Febr. 1962, 171—178.
[231] Ling R. F. A probability theory of cluster analysis, J. Amer. Statist. Assoc., 68, (1973), 159—164.
[232] Litofsky B. Utility of Automatic Classification Systems for Information Storage and Retrieval, Ph. D. Dissertation, University of Pennsylvania, cited in Dissertation Abstracts, Vol. 30, No. 7, (Jan., 1970), 3264-B.
[233] MacNaughton-Smith P. The classification of individuals by the possession of attributes associated with a criterion, Biometrics, 19, (1963), 364—366.
[234] MacNaughton-Smith P. Some statistical and other numerical techniques for classifying individuals. (home office res. rpt. no. 6) H. M. S. O., London, (1965).
[235] MacNaughton-Smith P. and Williams W. T., Dale M. B. and Mockett L. G. Dissimilarity analysis: A new technique of hierarchical division, Nature, Vol. 201, (1964), p. 426.
[236] MacQueen J. B. Some methods for classification and analysis of multivariate observations, Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, (1967), 281—297.
[237] MacQueen J. B. Some methods for classification and analysis of multivariate observations, Western Management Sci. Inst., University of California, working paper 96, (1966).
[238] McCammon R. B. and Wenninger G. The Dendograph, Computer Contribution 48, State Geological Survey, The University of Kansas, Lawrence, (1970), 28 pp.

[239] McCammon R. B. The dendograph: A new tool for correlation, Geological Society of America Bulletin, Vol. 79, (1968), 1163—1670.
[240] McQuitty L. L. Agreement analysis: Classifying persons by predominant patterns of responses, Brit. J. Statist. Psychol., Vol. 9, (1956), p. 5.
[241] McQuitty L. L. Single and multiple hierarchical classification by reciprocal pairs and rank order types, Educational and Psychological Measurement, 26, (1966), 253—265.
[242] McQuitty L. L. Agreement analysis: Classifying persons by predominant patterns of response, Brit. J. Statist. Psychol., 9, (1956), 5—16.
[243] McQuitty L. L. Capabilities and improvements of linkage analysis as a clustering method, Educational and Psychological Measurement, 24, (1964), 441—456.
[244] McQuitty L. L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevances. Educational and Psychological Measurement, 17, (1957), 207—229.
[245] McQuitty L. L. Hierarchical syndrome analysis, Educational and Psychological Measurement 20, (1960), 293—304.
[246] McQuitty L. L. Improved hierarchical syndrome analysis of discrete and continuous data, Educational and Psychological Measurement 26, (1966), 577—582.
[247] McQuitty L. L. Multiple hierarchical classification of institutions and persons with reference to union—measurement relations and psychological well-being, Educational and Psychological Measurement, 22, (1962), 513—531.
[248] McQuitty L. L. Rank order typal analysis, Educational and Psychological Measurement, 23, (1963), 55—61.
[249] McQuitty L. L. Typal Analysis, Educational and Psychological Measurement, 21, (1961), 677—696.
[250] McRae D. J. MIKCA: A FORTRAN IV Iterative K-means cluster analysis program, Behavioral Science, Vol. 16, No. 4, (1971), 423—424.
[251] Mahalonobis P. C. Analysis of race mixture in Bengal, J. Asiat. Soc. (India), Vol. 23, (1925), 301—310.
[252] Mahalanobis P. C. On the generalized distance in statistics, Proc. Natl. Inst. Sci. (India), Vol. 12, (1936), 49—55.
[253] Majone G. Distance-based cluster analysis and measurement scales, working paper no. 17, University of British Columbia, Vancouver, B. C., Canada, (Nov., 1968).
[254] Majone G. and Sanday P. R. On the Numerical Classification of Nominal Data, Report Number RR-118, Graduate School of Industrial Administration, Carnegie—Mellon University, (1968).
[255] Marill, T. and Green D. M. On the effectiveness of receptors in recognition systems, IEEE Trans. Information Theory, IT-9, Jan. 1963.
[256] Marriot F. H. C. A problem of optimum stratification, Biometrics, Vol. 26, (1970), 845—847.
[257] Marriot F. H. C. Practical problems in a method of cluster analysis, Biometrics, Vol. 27, (1971), 501—514.
[258] Mattson R. L. and Damman J. E. A technique for determining and coding sublasses in pattern recognition problems, IBM Journal, Jul 1965), 294—302.

[259] Matusita K. On the theory of statistical decision functions, Ann. Instit. Statist. Math. (Tokyo), Vol. 3, (1951), 17—35.
[260] Mayer L. S. Comment on the assignment of numbers to rank order categories, American Sociological Review, Vol. 35, No. 5, (1970), 916—917.
[261] Mayer L. S. A method of cluster analysis when there exist multiple indicators of a theoretic concept, Biometrics, Vol. 27, No. 1, (1971), 143—155.
[262] Mayer L. S. A note on treating original data as interval data, American Sociological Review, Vol. 36, No. 3, (1971), 518—519.
[263] Mayer L. S. A method of cluster analysis, paper presented at the joint statistical meetings, Fort Collins, Colorado, Aug. 23—26, 1971.
[264] Medgyessy Pal. Decomposition of Super—positions of Distribution Functions, Publishing House of Hungarian Academy of Science, Budapest, (Jun., 1961).
[265] Michener C. D. and Sokal R. R. A quantitative approach to a problem in classification, Evolution, Vol. 11, (Jun., 1957), 130—162.
[266] Michener C. D and Sokal R. R. A quantification of systematic relationships and phylogenetic trends, Proc. Xth International Congress of Entomogy, I, (1957), 409—415.
[267] Morishima H. and Oka H. The pattern of interspecific variations in the genus oryza: Its quantitatitive representation by statistical methods, Evolution, 14, (1960), 153—165.
[268] Morrison D. G. Measurement problems in cluster analysis, Management Sci., 13, (1967), 13-775-780.
[269] Morrison D. F. Multivariate Statistical Methods, McGraw-Hill Book Company, New York, (1967).
[270] Nagy G. State of the art of pattern recognition, Proc. IEEE, Vol. 56, (1968), 836—862.
[271] Nagy G. and Tolaba J. Nonsupervised crop classification through airborne MSS observations, IBM J. Res. and Dev., (Mar., 1972).
[272] Needham R. M. A method for using computers in information classification, Proc. I. F. I. P. Congress, 62, (1962), p. 284.
[273] Needham R. M. The theory of clumps, II, Report M. L., 139, Cambridge Language Research Unit, Cambridge, Eng., (Mar., 1961).
[274] Needham R. M. and Jones K. S. Keywords and clumps, J. Documentation, Vol. 20, (1964), p. 5.
[275] Nunnally J. The analysis of profile data, Psychol. Bull., Vol. 59, No. 4, (1962), 311—319.
[276] Okajima M., Stark L., Whipple G. and Yasui S. Computer pattern recognition techniques: Some results with real electrocardiographic data, IEEE Trans. on Bio-Medical Electronics, Vol. BME-10, No. 3, (Jul., 1963).
[277] Olds E. J. The 5% significance levels of sums of squares of rank differences and a correction, Ann. Math. Statist. 20, (1949).
[278] Ore O. Theory of graphs, Amer. Math. Soc., Providence, R. I., (1962).
Оре О. Теория графов. М., «Наука», 1968.

[279] Orr D. B. A new method for clustering jobs, J. Appl. Psychol. (1960), 44, 44—59.
[280] Parker-Rhodes A. F. Contributions to the theory of clumps, I. M. L. 138, Cambridge Language Research Unit, Cambridge England, (Mar., 1961).
[281] Parks J. M. Classification of mixed mode data by r-mode factor analysis and q-mode cluster analysis on distance functions, in A. J. Cole (ed.), Numerical Taxonomy, Academic Press, New York, (1969), 216—223.
[282] Parks J. M. FORTRAN IV Program for Q-Mode Cluster Analysis on Distance Function with Printed Dendogram, Computer Contribution 46, State Geological Survey, The University of Kansas, Lawrence, (1970).
[283] Parzen E. On estimation of a probability density function and mode, Ann. Math. Statist., Vol. 33, (1962), 1065—1076.
[284] Patrick E. A. and Hancock J. C. The non-supervised learning of probability spaces and recognition of patterns, Tech. Report, Purdue University, Lafayette, Ind., (1965).
[285] Patrick E. A. and Shen L. Y. L. Interactive use of problem knowledge for clustering and decision making, IEEE Trans. Computers, Vol. C-20, (Febr., 1971), 216—223.
[286] Pearson W. H. Estimation of a correlation measure from an uncertainty measure, Psychometrika, Vol. 31, No. 3, (1966), 421—433.
[287] Pearson E. S. and Hartley H. O. Biometric Tables for Statisticians, Cambridge University Press, Cambridge, England, (1954).
[288] Rand W. M. The Development of Objective Criteria for Evaluating Clustering Methods, Ph. D. Dissertation, UCLA, 138 pp. Cited in Dissertation Abstracts, Vol. 30, No. 11, (May, 1970), 4932-B.
[289] Rao C. R. The use and interpretation of principal components analysis in applied research, Sankhya, Series A, Vol. 26, (1964), 329—358.
[290] Rao C. R. The utilization of multiple measurements in problems of biological classification, J. Roy. Statist. Soc., Series B, 10, (1948), 159—203.
[291] Rao M. R. Cluster Analysis and Mathematical Programming, Journal of the American Statistical Association, J. Amer. Statist. Assoc., Vol. 66, (1971), 622—626.
[292] Reiter S. and Sherman G. Discrete optimizing, J. S. I. A. M., 13, (1965), 864—889.
[293] Rescigno A. and Maccacaro G. A. The information content of biological classifications, in C. Cherry (ed.), Information Theory, 4 th London Symposium, Butterworths, London, (1961), 437—446.
[294] Rogers D. J. and Tanimoto T. T. A computer program for classifying plants, Science, Vol. 132, (Oct. 21, 1960), 1115—1118.
[295] Rohlf F. J. Adaptive hierarchical clustering schemes, Systematic Zoology, Vol. 19, No. 1, (1970), 58—83.
[296] Rohlf F. J. and Sokal R. R. Coefficient of correlation and distance in numerical taxonomy, Kansas University Sci. Bull., 45, (1965), 3—27.

[297] Rose M. J. Classification of a set of elements, Comput. J., Vol. 7, (1964), p. 208.
[298] Rosenblatt M. Remarks on some nonparametric estimates of a density function, AMS 27, (1956), 832—837.
[299] Ross G. J. S. Algorithms AS 13—15. Appl. Statist., 18, (1969), 103—110.
[300] Rota G. The number of partions of a set, Amer. Math. Monthly, Vol. 71, No. 5, 498—504.
[301] Rubin J. Optimal classification into groups: An approach for solving the taxonomy problem, J. Theoret. Biol., (1967), 15, 103—144.
[302] Ruspini E. H. A new approach to clustering, Information and Control, Vol. 15, (1969), 22—32.
[303] Russell P. F. and Rao T. R. On habitat and association of species of anopheline larvae in South-eastern Madras, J. Malor Inst. India 3, (1940), 153—178.
[304] Sahler W. A survey of distribution — free statistics based on distances between distribution functions, Metrika, Vol. 13, (1968), 149—169.
[305] Sammon J. W. Interactive pattern analysis and classification, IEEE Trans. on Computers, Vol. C-19, No. 7, (Jul., 1970), 594—610.
[306] Sammon J. W. Jr. On-Line Pattern Analysis and Recognition System (OLPARS), report number RADC-TR-68-263, Rome Air Development Center, Griffiss Air Force Base, New York, 72 pp., (1968).
[307] Samuel E. and Bachi R. Measures of Distances of Distribution Functions and Some Applications, Metron, Vol. 23, (Dec., 1964), 83—122.
[308] Sandon F. The means of sections from a normal distribution, Brit. J. Statist. Psychol., Vol. XIV, Part II, (1961), 117—121.
[309] Sawrey W. L., Keller L. and Conger J. J. An objective method of grouping profiles by distance functions and its relation to factor analysis, Educational and Psychological Measurement, Vol. 20, No. 4, (1960).
[310] Schnell P. Eine methode zur auffindung von gruppen, Biom. Z., Vol. 6, (1964), 47—48.
[311] Schweitzer S. and Schweitzer D. G. Comment on the Pearson r in random number and precise functional scale transformations, Amer. Sociol. Rev., Vol. 36, No. 3, (1971), 517—518.
[312] Scott A. J. and Symons M. J. Clustering methods based on likelihood ratio criteria, Biometrics, Vol. 27, No. 2, (1971), 387—398.
[313] Scott A. J. and Symons M. J. On the Edwards and Cavalli-Sforza method of cluster analysis, Biometrics, Vol. 27, No. 1, (1971), 217—219.
[314] Sebestyen George S. Automatic off-line multivariate data analysis, Proc. Fall Joint Comput. Conf., Spartan Books, (Nov., 1966), 685—694.
[315] Sebestyen G. S. Pattern recognition by an adaptive process of sample set construciton, IRE Trans. on Info. Theory, Vol. IT-8, (Sept. 1962).
[316] Sebestyen G. S. Recognition of membership in classes, IRE Trans. Info. Theor., (1961), IT-7 (1), 48—50.

[317] Sebestyen G. and Edie J. An algorithm for nonparametric pattern recognition, IEEE Trans. Electronic Computers, Vol. EC-15, No. 6, (1966), 908—915.
[318] Sebestyen G. S. and Edie J. Pattern recognition research, Air Force Cambridge Res. Lab. Report 64—821, Bedford, Mass., (Jun., 14, 1964).
[319] Shepherd M. J. and Willmott A. J. Cluster Analysis on the atlas computer, Comput. J., Vol. 11, (1968), 57—62.
[320] Shepard R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, Psychometrika, 27, (1962), 125—129, 219—246.
[321] Shepard R. N. and Carroll J. D. Parametric representation of nonlinear data structures, in P. R. Krishnaiah, Multivariate Analysis, Academic Press, New York, (1966).
[322] Sibson R. A model for taxonomy II, Math. Biosci., 6, (1970), 405—430.
[323] Sibson R. Some observations of a paper by Lance and Williams, Comput. J., 14, (1971), 156—157.
[324] Silverman J. A computer technique for clustering tasks, Technical Bull. STB 66—23, San. Diego, Calif.: U. S. Naval Personnel Research Activity, (Apr., 1966).
[325] Singleton R. C. Minimum squared-error clustering, Unpublished Internal Communication at Stanford Research Institute, Menlo Park, California, (1967).
[326] Smirnov N. V. On the estimation of the discrepancy between empirical curves of distribution for two independent samples, Bull. Math. Univ. Moscow, Vol. 2, (1939), 3—14.
Смирнов Н. В. Оценка расхождения между эмпирическими кривыми распределениями в двух независимых выборках. Бюлл. МГУ, т. II, вып. 7, 1939.
[327] Smith J. W. The analysis of multiple signal data, IEEE Trans. on Information Theory, Vol. IT-10, No. 3, (Jul., 1964).
[328] Sneath P. H. A. The application of computers to taxonomy, J. General Microbiology 17, (1957), 201—226.
[329] Sneath P. H. A. A comparison of different clustering methods as applied to randomly spaced points, Classification Soc. Bull., 1, (1966), 2—18.
[330] Sneath P. H. A. A method for curve seeking from scattered points, Comput. J., Vol. 8, (1966), p. 383.
[331] Sneath P. H. A. Evaluation of clustering methods, in A. J. Cole (ed.), Numerical Taxonomy, Academic Press, London and New York, (1969), 257—267.
[332] Sonquist J. A. and Morgan J. N. The Detection of Interaction Effects, Survey Research Center, Institute for Social Research, University of Michigan, Ann. Arbor, (1964).
[333] Sokal R. R. Distances as a measure of taxonomic similarity, Systematic Zoology, 10, (1961), 70—79.
[334] Sokal R. R. and Michener C. D. A statistical method for evaluating systematic relationships, University of Kansas Sci. Bull., (Mar., 20, 1958), 1409—1438.
[335] Sokal R. R. and Rohlf F. J. The comparison of dendrograms by objective methods, Taxon, Vol. 11, (1962), 33—40.
[336] Sokal R. R. and Sneath P. H. A. Principles of Numerical Taxonomy, San Francisco: W. H. Freeman and Company, (1963).

[337] Solomon H. Numerical Taxonomy, Technical Report Number 167, Department of Statistics, Stanford University, 44 p., (1970).
[338] Sorenson T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application fo analyses of the vegetation on Danish commons, Biol. Skr. 5, (1968), 1—34.
[339] Spearman C. Correlations of sums and differences, Brit. J. Psychol. 5, (1913), 417—426.
[340] Speilberg K. On the fixed charge transporatation problem, Proc. 19 th Natl. Conf., Association for Computing Machinery, 211 East 43 rd Street, New York, (Aug., 1964).
[341] Spilker J. J. Jr., Luby D. D. and Lawhorn R. D. Progress report — adaptive binary waveform detection, Tech. Report 75, Communication Sciences Department, Philco Corp., Palo Alto, Calif., (Dec., 1963).
[342] Srikantan K. S. Canonical association between nominal measurements, J. Amer. Statist. Assoc., Vol. 65, No. 329, (1970), 284—292.
[343] Stanat Donald F. Nonsupervised pattern recognition through the decomposition of probability functions, Tech. report, Sensory Intelligence Lab, Dept. of Psychology, University of Michigan, (Apr., 1966), 1—55.
[344] Stark, Lawrence, Okajima, Mitsuharu and Whipple Gerald H. Computer pattern recognition techniques: Electrocardiographic diagnosis, Comm. ACM, 6, No. 10, (Oct., 1962), 527—532.
[345] Steinbuch K. and Piske U. A. W. Learning matrices and their applications, IEEE Trans. on Electronic Computers, Vol. EC-12, No. 6, (Dec., 1963).
[346] Stephenson W. Correlating persons instead of tests, Character and Personality, Vol. 4, No. 1, (1935), 17—24.
[347] Stephenson W. The inverted factor technique, Brit. J. Psychol., Vol. 26, No. 4, (1938), 344—361.
[348] Stephenson W. The Study of Behavior, University of Chicago Press, Chicago, Ill., (1953).
[349] Stewart D. and Love W. A general cononical correlation index, Psychol. Bull., Vol. 70, No. 3, (1968), 160—163.
[350] Stiles H. E. The association factor in information retrieval, Comm. ACM, Vol. 8, No. 1, (1961), 271—279.
[351] Stringer P. Cluster analysis of nonverbal judgements of facial expressions, Brit. J. Math. Statist. Psychol., (1967), 20 (1), 71—79.
[352] Stuart A. The correlation between variate values and ranks in samples from a continuous distribution, Brit. J. Statist. Psychol., Vol. VII, Part I, (1954), 37—44.
[353] Swain P. H. and Fu K. S. Nonparametric and linguistic approaches to pattern recognition, LARS Information Note 051970, Purdue University, Lafayette, Indiana, (Jun., 1970).
[354] Switzer P. Statistical techniques in clustering and pattern recognition, Department of Statistics, Stanford Univ., TR139.
[355] Tanimoto T. T. and Loomis R. G. A taxonomy program for the IBM 704, New York, International Business Machines Corporation (Data Systems Division, Mathematics and Applications Department), (1960), (M & A-6, the IBM Taxonomy Application.)

[356] Tatsuoka M. M. The Relationship Between Canonical Correlation and Discriminant Analysis, and a Proposal for Utilizing Quantitative Data in Discriminant Analysis, Educational Research Corporation, Cambridge, Mass. (1955).
[357] Thomas L. L. A cluster analysis of office operations, J. Appl. Psychol., (1952), 36, 238—242.
[358] Thompson W. A. Jr. The problem of negative estimates of variance components, AMS, 33, (1962), 273—289.
[359] Thorndike R. L. Who belongs in the family, Psychometrika, 18 (1953), 267—276.
[360] Torgerson W. S. Multidimensional scaling of similarity, Psychometrika, Vol. 30, No. 4, (1965), 379—393.
[361] Tryon R. C. Cluster Analysis, Ann Arbor: Edwards Bros., (1939).
[362] Tryon R. C. Cluster analysis, Psychometrika, Vol. 22, No. 3, (Sept., 1957), 241—260.
[363] Tryon R. C. Communality of a variable: Reformulation by cluster analysis, Psychometrika, 22, (1957), 241—260.
[364] Tryon R. C. Comparative cluster analysis, Psychol. Bull., 36, (1939), 645—646.
[365] Tryon R. C. Commulative communality cluster analysis, Educ. Psychol. Measmt., (1958), 18, 3—35.
[366] Tryon R. C. Domain sampling formulation of cluster and factor analysis, Psychometrika, Vol. 24, No. 2, (Jun., 1959), 113—135.
[367] Tryon R. C. General dimensions of individual differences: Cluster analysis versus multiple factor analysis, Educ. Psychol. Measmt., (1958), 18, 477—495.
[368] Tryon R. C. Identification of social areas by cluster analysis, California: University of California Press, (1955).
[369] Tryon R. C. Improved cluster — orthometrics, Psychol. Bull., 36, (1939), p. 529.
[370] Tryon R. C. and Bailey D. E. The BC TRY computer system of cluster and factor analysis, Multivar. Behav. Res., (1966), 1, 95—111.
[371] Tryon R. C. and Bailey D. E. Cluster Analysis, McGraw-Hill Book Company, New York, (1970).
[372] Tucker Ledyard R. Cluster analysis and the search for structure underlying individual differences in psychological phenomena, Conference on Cluster Analysis of Multivariate Data, New Orleans, La., (Dec., 1966), 10.01—10.17.
[373] Turner B. J. Cluster analysis of MSS remote sensor data, presented by Conference on Earth Resources Observation and Information Analysis Systems, Tullahoma, Tenn., (Mar., 1972).
[374] Turner R. D. First — order experimental concept formation, Biological Prototypes and Synthetic Systems, E. E. Bernard and M. Kare, eds., Bionics Symposium 2, Ithaca, N. Y., Plenum Co., 1961.
[375] Urbakh V. Yu. A discriminate method of clustering, J. Multivariate Analysis, Vol. 2, (1972), 249—260.
[376] Urbakh V. Yu. On decomposition of statistical distributions deviating from normal into two normal distributions, Biofizika (Moskow), Vol. 6, (1961), 265—271.
[377] Van Rijsbergen C. J. A clustering algorithm, Comput J., 13, (1970), 113—115, (algorithm 47).

[378] Van Rijsbergen C. J. A fast hierarchic clustering algorithm, Comput. J., 13, (1970), 324—326.
[379] Vargo L. G. Comment on «The Assignment of Numbers to Rank Order Categories», Amer. Sociol. Rev., Vol. 36, No. 3, (1971), 516—517.
[380] Vinod H. D. Integer programming and theory of grouping, JASA, (Jun., 1969), 506—519.
[381] Von Mises R. Wahrscheinlichkeitsrechnung, Leipzig — Wein, (1931).
[382] Wacker A. G. and Langrebe D. A. Boundaries in MSS imaging by clustering, Proc. 9 th IEEE Symposium on Adaptive Processes, (Dec., 1970).
[383] Wacker A. G. and Langrebe D. A. The minimum distance approach to classification, The Laboratory for Applications of Remote Sensing Information Note 100771, Purdue University, Lafayette, Indiana, (Oct. 1971).
[384] Wallace C. S. and Boulton B. M. An information measure for classification, Comput. J., Vol. 11, (1968), p. 185.
[385] Watson L., Williams W. T. and Lance G. N. Angiosperm Taxonomy: A comparative study of some novel numerical techniques, J. Linn. Soc., Vol. 59, (1966), p. 491.
[386] Ward J. H. Jr. Hierarchical grouping to maximize payoff, WADD-TN-61-29, Lackland AFB, Texas: Personnel Laboratory, Wright Air Development Division, (Mar., 1961).
[387] Ward J. H. Jr. Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc., Vol. 58, No. 301, (1963), 236—244.
[388] Ward J. H. Jr., Hall K. and Buchhorn J. PERSUB Reference Manual, Report No. PRL-TR-67-3 (II), Personnel Research Laboratory, Lackland AFB, Texas, 60 pp., (1967).
[389] Ward J. H. Jr. and Hook M. E. Applications of an hierarchical grouping procedure to a problem of grouping profiles, Educational and Psychological Measurement, Vol. 23, No. 1, (1963), 69—82.
[390] Wherry R. J. Jr. and Lane N. E. The K-Coefficient, A Pearson — Type Substitute for the Contingency Coefficient, Report No. NSAM-929, Naval School of Aviation Medicine, Pensacola, Florida, 20 pp., (1965).
[391] Wilks S. S. Mathematical Statistics, John Wiley and Sons, Inc., New York, (1962).
Уилкс С. Математическая статистика. М., «Наука», 1967.
[392] Williams E. J. Use of scores for the analysis of association in contingency tables, Biometrika, Vol. 39, Part 3/4, (1952), 274—289.
[393] Williams W. T., Dale M. B. and Macnaughton-Smith P. An objective method of weighting in similarity analysis, Nature, Vol. 201, (1964), p. 426.
[394] Williams W. T. and Lambert J. M. Multivariate methods in plant ecology I. Association analysis in plant communities, J. Ecology, Vol. 47, No. 1, (1959), 83—101.
[395] Williams W. T., Lambert J. M. and Lance G. N. Multivariate Methods in Plant Ecology, J. Ecology, Vol. 54, p. 427.
[396] Wishart D. Mode analysis: A generalization of nearest neighbor which reduces chaining effects, in A. J. Cole (ed.), Numerical Taxonomy, Academic Press, New York, (1969), 282—319.

[397] Wishart D. An algorithm for hierarchical classifications, Biometrics, Vol. 22, No. 1, (1969), 165—170.
[398] Wishart D. FORTRAN II Programs for 8 Methods of Cluster Analysis (CLUSTAN I), Computer Contribution 38, State Geological Survey, The University of Kansas, Lawrence, (1969).
[399] Wishart D. A fortran II program for numerical classification, St. Andrew's University, Scotland, (1968).
[400] Wishart D. Numerical classification method for deriving natural classes, Nature, London, (1969), 221, 97—98.
[401] Wolf D. E. PROMENADE: Complete Listing of PROMENADE Programs, Appendix 9d to RADC-TR-68-572, Stanford Research Institute, Menlo Park, Calif., 465 pp., (1968).
[402] Wolfe J. H. A computer program for the maximum likelihood analysis of types, Tech. Bulletin, 65—15, U. S. Naval Personnel Research Activity, San Diego, Calif., (May, 1965).
[403] Wolfe John H. NORMIX-Computational methods for estimating the parameters of multivariate normal mixtures of distributions, Tech. Report, U. S. Naval Personnel Research Activity, San Diego, Calif., (Aug., 1967), 1—31.
[404] Wolfe J. H. Pattern clustering by multivariate mixture analysis, Multivariate Behavioral Research, Vol. 5, No. 3, (1970), 329—350.
[405] Young G. Factor analysis and the index of clustering, Psychometrika, Vol. 4, No. 3, (Sept., 1939).
[406] Yule G. U. On measuring associations between attributes, J. Roy. Statist. Soc., Vol. 75, (1912), 579—642.
[407] Zadeh L. A. Fuzzy sets, Information and Control, Vol. 8, (1965), 338—353.
[408] Zahn C. T. Graph — theoretical methods for detecting and describing gestalt clusters, IEEE Trans. on Computers, Vol. C-20, No. 1, (1971), 68—86.
[409] Remote multispectral sensing in agriculture, Laboratory for Applications of Remote Sensing, Purdue University, Lafayette, Ind., Annual Report, Vol. 4, Research Bulletin 873, (Dec., 1970).

Литература, добавленная при переводе

С. А. Айвазян, З. И. Бежаева, О. В. Староверов. Классификация многомерных наблюдений. М., «Статистика», 1974.
Многомерный статистический анализ в социально-экономических исследованиях. М., «Наука», 1974, гл. II.

ОГЛАВЛЕНИЕ

О методологических принципах и многомерном анализе (вместо предисловия) 5
Предисловие 13
Глава 1. Проблема кластерного анализа. Основные идеи . 14
1.1 Основные обозначения и определения 14
1.2 Задача кластерного анализа 15
1.3 Функции расстояния 16
1.4 Меры сходства 18
1.5 Расстояние между кластерами и их сходство . . 23
1.6 Кластерные методы, основанные на евклидовой метрике 27
1.7 Алгоритм последовательной кластеризации . . . 35
1.8 Другие вопросы кластерного анализа 39
Глава 2. Кластеризация полным перебором 41
2.1 Введение 41
2.2 Число разбиений n объектов на m непустых подмножеств 41
2.3 Рекурсивное соотношение между числами Стирлинга второго рода 47
2.4 Вычислительные аспекты полного перебора 49
Глава 3. Математическое программирование и кластерный анализ 50
3.1 Применение динамического программирования к кластерному анализу 50
3.2 Модель динамического программирования Дженсена 57
3.3 Применение целочисленного программирования в кластерном анализе 64
Глава 4. Представления матриц сходств 72
4.1 Дендограммы 72
4.2 Сравнение дендограмм и матриц сходства 77
4.3 Основные определения 78
4.4 Деревья 79
4.5 Локальные операции на деревьях 84
Глава 5. Кластеризация на основе оценивания функции плотности 87
5.1 Модальный анализ 87
5.2 Оценивание функции плотности вероятности . . . 89
5.3 Кластеризация на основе оценивания функции плотности 93
5.4 Замечания 95

Глава 6. Приложения 96
6.1. Приложение к регистрации отдаленных объектов . 96
6.2 Применение метода оценивания функции плотности для данных Фишера по ирису [40] 99
Глава 7. Исторические замечания 100
Литература 105

Дюран Б. и Оделл П.

КЛАСТЕРНЫЙ АНАЛИЗ

Редактор *З. А. Сумник*. Мл. редактор *О. В. Степанченко*
Техн. редактор *В. А. Чуракова*. Корректор *Г. А. Башарина*
Худ. редактор *Н. А. Володина*
Обложка художника *Г. Г. Васильевой*

ИБ № 345

Сдано в набор 12/X 1976 г. Подписано к печати 21/I 1977 г. Формат бумаги 84×108$^{1}/_{32}$. Бумага № 3. Объем 4,0 печ. л. Уч.-изд. л. 7,05. Усл. п. л. 6,72. Тираж 11 000 экз. (Тематич. план 1977 г. № 40). Заказ № 6796. Цена 42 коп.

Издательство «Статистика», Москва, ул. Кирова, 39.

Областная типография управления издательств, полиграфии и книжной торговли Ивановского облисполкома, г. Иваново-8; ул. Типографская, 6.