**1. Введение в кластерный анализ**

Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster – гроздь, скопление. Впервые в 1939 был определен предмет кластерного анализа и сделано его описание исследователем Трионом (Tryon). Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней. Методы кластерного анализа можно применять в самых различных случаях, даже в тех случаях, когда речь идет о простой группировке, в которой все сводится к образованию групп по количественному сходству.

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы социально-экономической информации, делать их компактными и наглядными. Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, характеризующих экономическое развитие (например, общехозяйственной и товарной конъюнктуры). Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл здесь может давать информацию, которая способна сильно изменить направленность и подходы дальнейшего применения кластерного анализа. Этот процесс можно представить системой с обратной связью.

В задачах социально-экономического прогнозирования весьма перспективно сочетание кластерного анализа с другими количественными методами (например, с регрессионным анализом).

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения: В частности, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов за счет замены их характеристиками обобщенных значений параметров кластера. При проведении классификации объектов игнорируется очень часто возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров.

В кластерном анализе считается, что:

а) выбранные характеристики допускают в принципе желательное разбиение на кластеры;

б) единицы измерения (масштаб) выбраны правильно.

Выбор масштаба играет большую роль. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклонение, так что дисперсия оказывается равной единице.

*Задача кластерного анализа* заключается в том, чтобы на основании данных, содержащихся во множестве Х, разбить множество объектов G на m (m – целое) кластеров (подмножеств) Q1, Q2, …, Qm, так, чтобы каждый объект Gj принадлежал одному и только одному подмножеству разбиения. А объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были разнородными.

*Решением задачи кластерного анализа* являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

где xj - представляет собой измерения j-го объекта.

Кластер имеет следующие *математические характеристики*: *центр, радиус, среднеквадратическое отклонение, размер кластера*.

*Центр кластера* - это среднее геометрическое место точек в пространстве переменных.

*Радиус кластера* - максимальное расстояние точек от центра кластера. Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют *спорными*.

*Спорный объект* - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

*Размер кластера* может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до *центра кластера* меньше *радиуса кластера*. Если это условие выполняется для двух и более кластеров, объект является спорным.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение пула (совокупности) объектов на кластеры. Мы уже упоминали о сравнимости шкал, это и есть второе предположение - правильность выбора масштаба или единиц измерения признаков.

Выбор масштаба в кластерном анализе имеет большое значение. Рассмотрим пример. Представим себе, что данные признака х в наборе данных А на два порядка больше данных признака у: значения переменной х находятся в диапазоне от 100 до 700, а значения переменной у - в диапазоне от 0 до 1.

Тогда, при расчёте величины расстояния между точками, отражающими положение объектов в пространстве их свойств, переменная, имеющая большие значения, т.е. переменная х, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной у. Таким образом из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками.

Эта проблема решается при помощи предварительной *стандартизации* переменных. *Стандартизация* (standardization) или нормирование (normalization) приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через отношение этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных:

где - соответственно среднее и среднеквадратическое отклонение *x*; xmax, xmin - наибольшее и наименьшее значение *x*.

Наряду со *стандартизацией* переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы отражал значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки, полученные в ходе опроса экспертов - специалистов предметной области. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных.

В ходе экспериментов возможно сравнение результатов, полученных с учётом экспертных оценок и без них, и выбор лучшего из них.

**1.1 Измерение близости объектов**

Проблема измерения близости объектов неизбежно возникает при любых трактовках кластеров и различных методах классификации.

Отметим основные трудности, возникающие при этом: неоднозначность выбора способа нормировки и определения расстояния между объектами. Рассмотрим результаты небольшого обследования. Студенты группы записывают свои данные (вес, рост), оформляют в таблицу и строят по ним корреляционное поле. Масштабы по осям выбираются произвольно (рисунок 1.1).

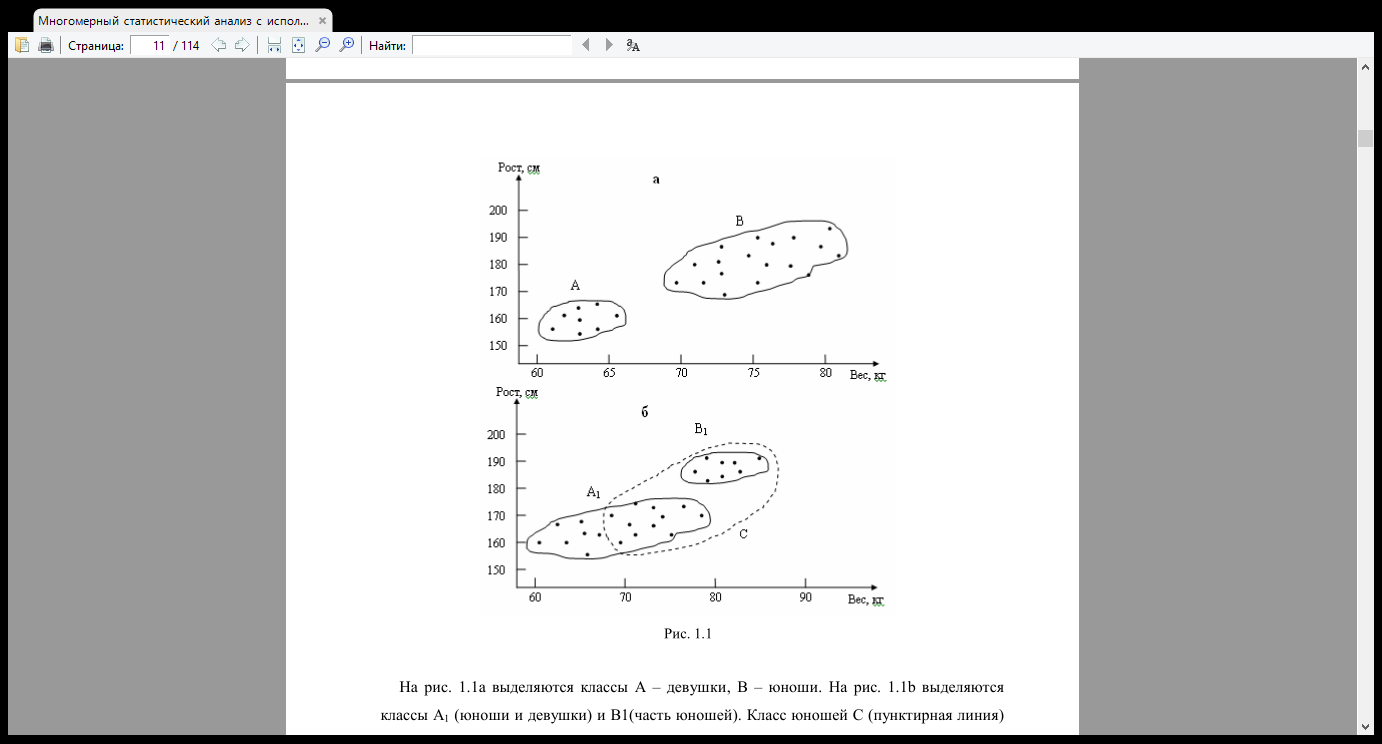


Рисунок 1.1

На рисунке 1.1а выделяются классы A – девушки, B – юноши. На рис. 1.1б выделяются классы A1 (юноши и девушки) и B1(часть юношей). Класс юношей C (пунктирная линия) на рис. 1.1б не выделит, поскольку расстояния между ближайшими объектами классов A1 и B1 существенно больше, чем внутренние расстояния в A1, юноши из A почти никакими алгоритмами к B1 не присоединяются.

Однако определить расстояние между объектами в данном случае нельзя, поскольку признаки измерены в разных единицах измерения. Требуется нормировка показателей, переводящая их в безразмерные величины: тогда измерение близости объектов становится оправданным.

В кластерном анализе для количественной оценки сходства вводится понятие *метрики*. Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается *k* признаками, то он может быть представлен как точка в *k*-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние.

*Расстоянием (метрикой)* между объектами в пространстве параметров называется такая величина *dab*, которая удовлетворяет аксиомам:

A1. *dab* > 0, dab = 0,

A2. *dab* = dba,

A3. *dab* + dbc ≥ dac.

*Мерой близости (сходства)* обычно называется величина µ*ab*, имеющая предел и возрастающая с возрастанием близости объектов.

B1. µ*ab* непрерывна,

B2. µ*ab* = µ*ba*,

B3. 1 ≤ µ*ab* ≤ 0.

Существует возможность простого перехода от расстояний к мерам близости:

**1.2 Характеристики близости объектов**

Объединение или метод древовидной кластеризации используется при формировании кластеров несходства или расстояния между объектами. Эти расстояния могут определяться в одномерном или многомерном пространстве. Например, если вы должны кластеризовать типы еды в кафе, то можете принять во внимание количество содержащихся в ней калорий, цену, субъективную оценку вкуса и т.д. Наиболее прямой путь вычисления расстояний между объектами в многомерном пространстве состоит в вычислении евклидовых расстояний. Если вы имеете двух- или трёхмерное пространство, то эта мера является реальным геометрическим расстоянием между объектами в пространстве (как будто расстояния между объектами измерены рулеткой). Однако алгоритм объединения не "заботится" о том, являются ли "предоставленные" для этого расстояния настоящими или некоторыми другими производными мерами расстояния, что более значимо для исследователя; и задачей исследователей является подобрать правильный метод для специфических применений.

Рассмотрим основные способы определения близости между объектами (Таблица 1.1)

Таблица 1.1

|  |  |
| --- | --- |
| Показатели | Формулы |
| Для количественных шкал | |
| Линейное расстояние |  |
| Евклидово расстояние |  |
| Квадрат евклидова расстояния |  |
| Обобщенное степенное расстояние Минковского |  |
| Расстояние Чебышева |  |
| Расстояние городских кварталов  (Манхэттенское расстояние) |  |

*Евклидово расстояние* является самой популярной метрикой в кластерном анализе. Оно попросту является геометрическим расстоянием в многомерном пространстве. Геометрически оно лучше всего объединяет объекты в шарообразных скоплениях.

*Квадрат евклидова расстояния*. Для придания больших весов более отдаленным друг от друга объектам можем воспользоваться квадратом *евклидова расстояния* путем возведения в квадрат стандартного *евклидова расстояния*.

*Обобщенное степенное* расстояние представляет только математический интерес как универсальная метрика.

*Расстояние Чебышева*. Это расстояние стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

*Манхэттенское расстояние* (расстояние городских кварталов), также называемое "хэмминговым" или "сити-блок" расстоянием. Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат.

**1.3 Методы кластерного анализа**

Методы кластерного анализа можно разделить на две группы: иерархические и неиерархические. Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

**1.3.1 Иерархические алгоритмы**

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

*Иерархические агломеративные методы (Agglomerative Nesting, AGNES)*

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

*Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA)*

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде дендрограммы показан на рисунке 1.2.

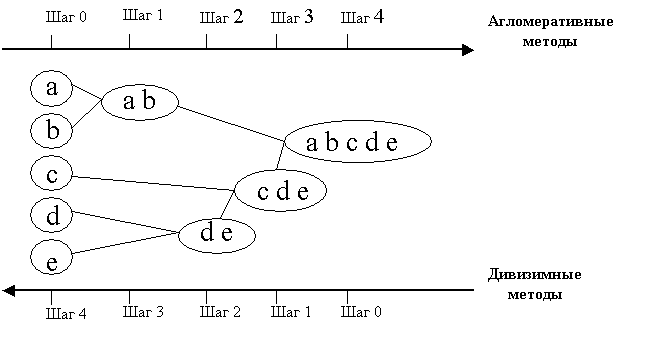


Рисунок 1.2

Дендрограмма агломеративных и дивизимных методов Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных. Преимуществом иерархических методов кластеризации является их наглядность.

Иерархические алгоритмы связаны с построением дендрограмм (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров. Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

Существует много способов построения дендограмм. В дендограмме объекты могут располагаться вертикально или горизонтально. Пример вертикальной дендрограммы приведен на рисунке 1.3.

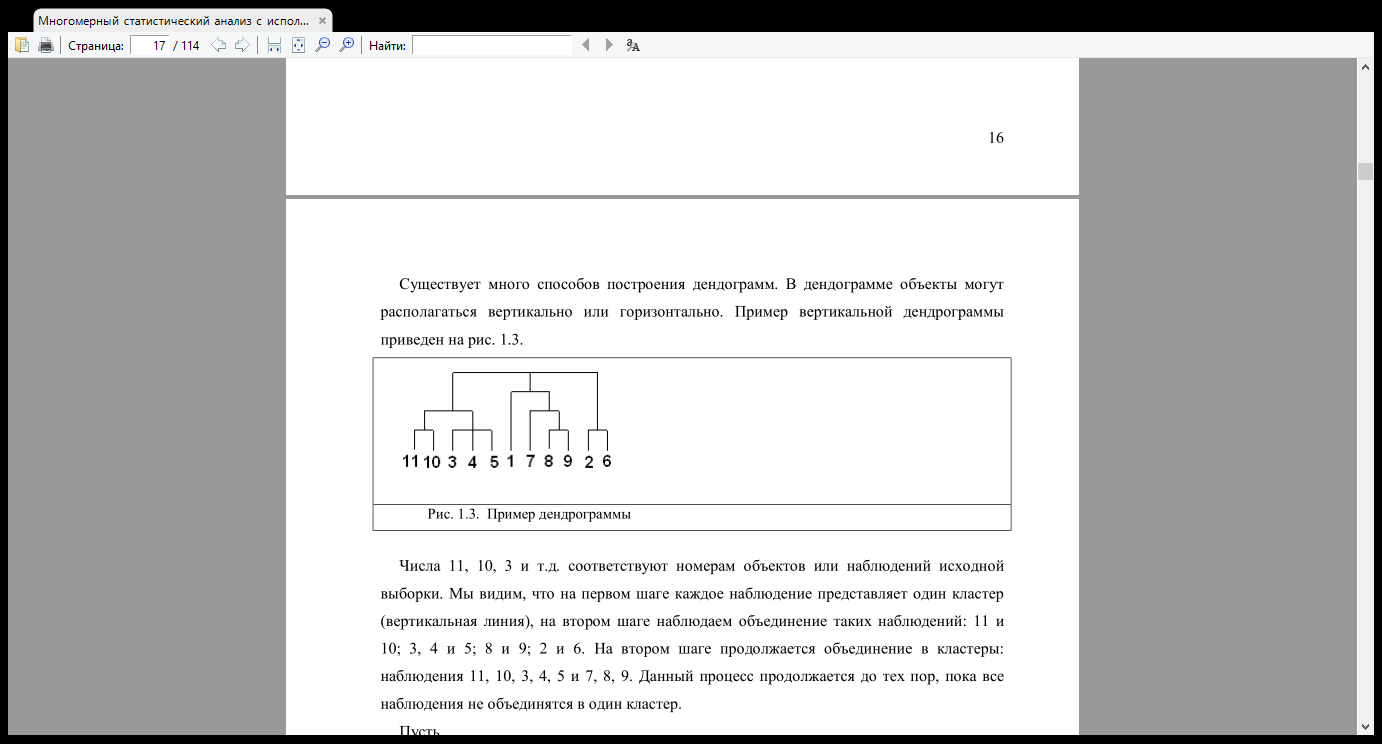


Рис. 1.3 - Пример дендрограммы

Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. Мы видим, что на первом шаге каждое наблюдение представляет один кластер (вертикальная линия), на втором шаге наблюдаем объединение таких наблюдений: 11 и 10; 3, 4 и 5; 8 и 9; 2 и 6. На втором шаге продолжается объединение в кластеры: наблюдения 11, 10, 3, 4, 5 и 7, 8, 9. Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Пусть Ki - i-я группа (класс, кластер), состоящая из n объектов;

- среднее арифметическое векторных наблюдений Ki группы, т.е. «центр тяжести» i-й группы;

– расстояние между группами Ki и Kj.

*Обобщённая агломеративная процедура*. На первом шаге каждый объект считается отдельным кластером. На следующем шаге объединяются два ближайших объекта, которые образуют новый класс, определяются расстояния от этого класса до всех остальных объектов, и размерность матрицы расстояний D сокращается на единицу. На p-ом шаге повторяется та же процедура на матрице D(n- p)(n- p) , пока все объекты не объединятся в один класс.

Если сразу несколько объектов (классов) имеют минимальное расстояние, то возможны две стратегии: выбрать одну случайную пару или объединить сразу все пары. Первый способ является классическим и реализован во всех процедурах (иногда его называют восходящей иерархической классификацией). Второй способ называют методом ближайших соседей (не путать с алгоритмом “*Ближайшего соседа*”) и используют реже.

Результаты работы всех иерархических процедур обычно оформляют в виде так называемой *дендрограммы* (рисунки 1.4.1 – 1.4.3). В дендрограмме номера объектов располагаются по горизонтали, а по вертикали - результаты кластеризации.

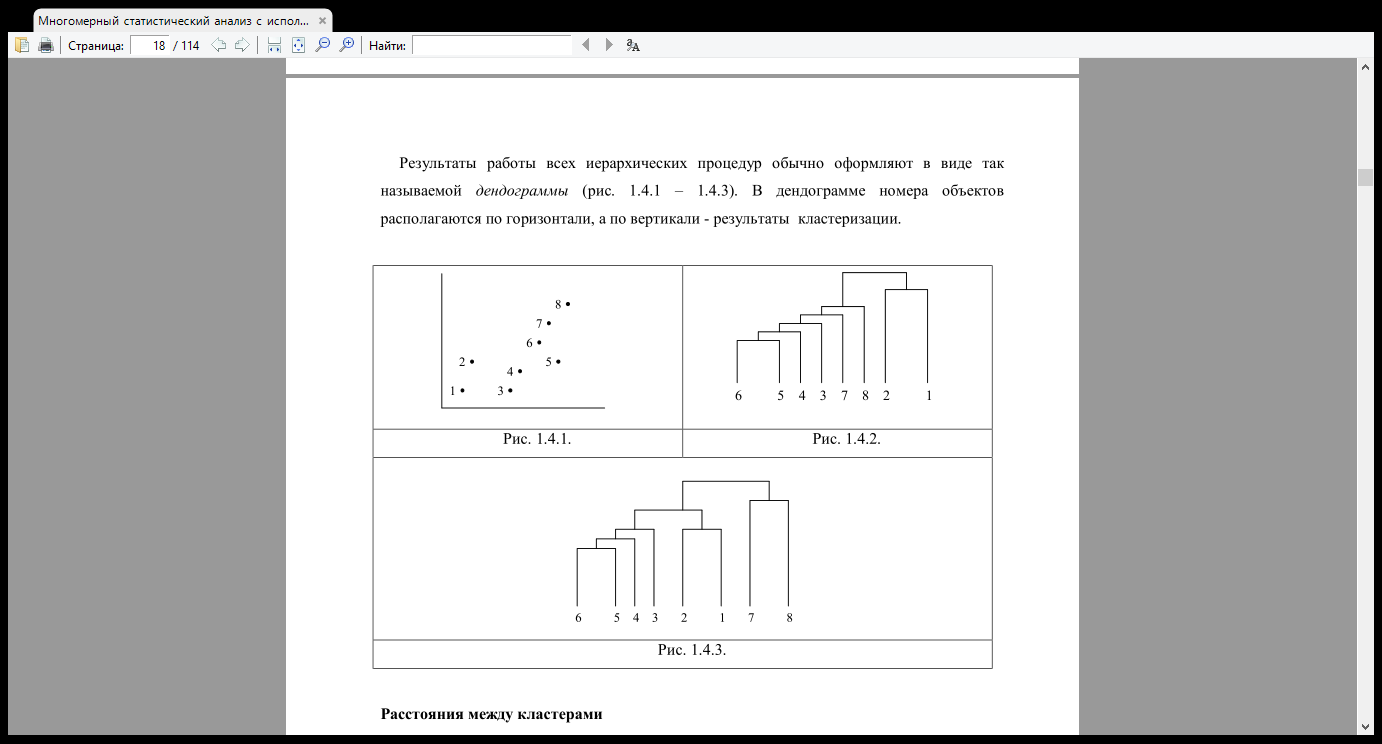


Рисунок 1.4.1

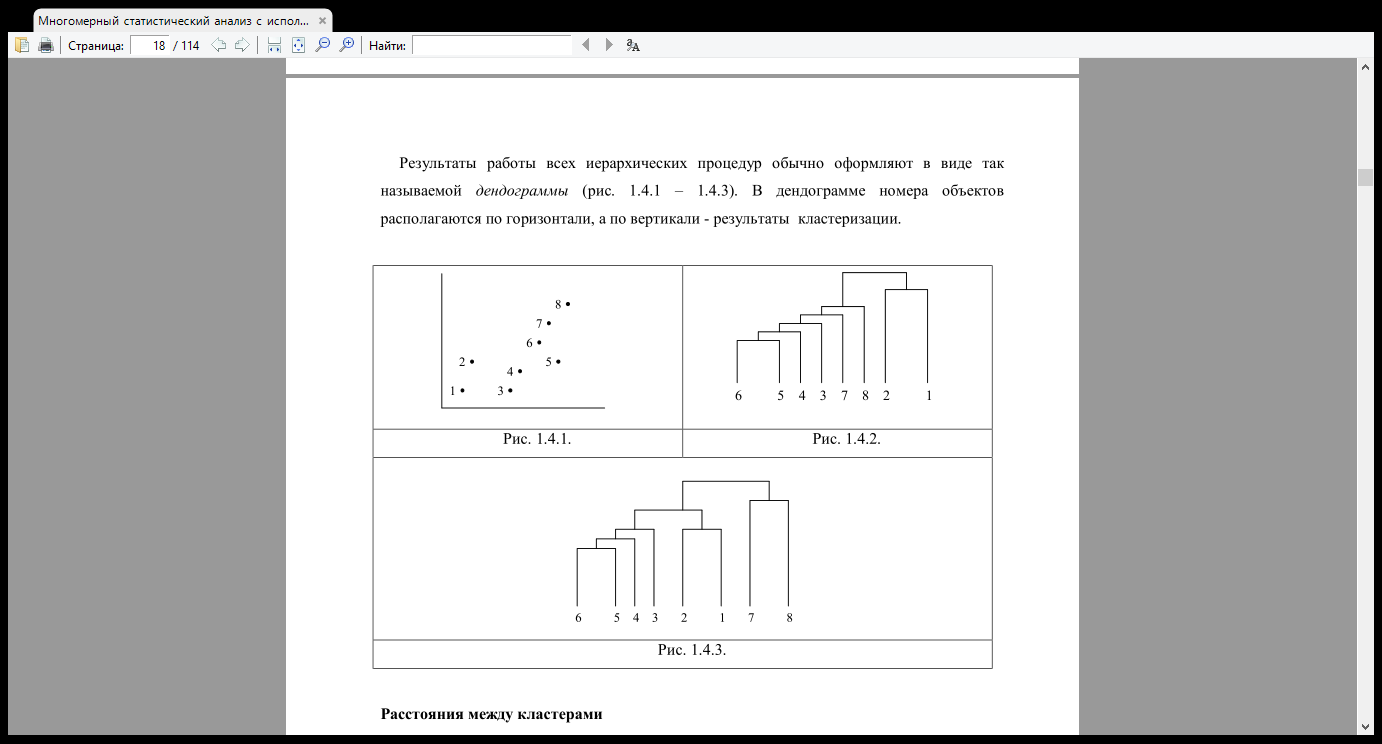


Рисунок 1.4.2

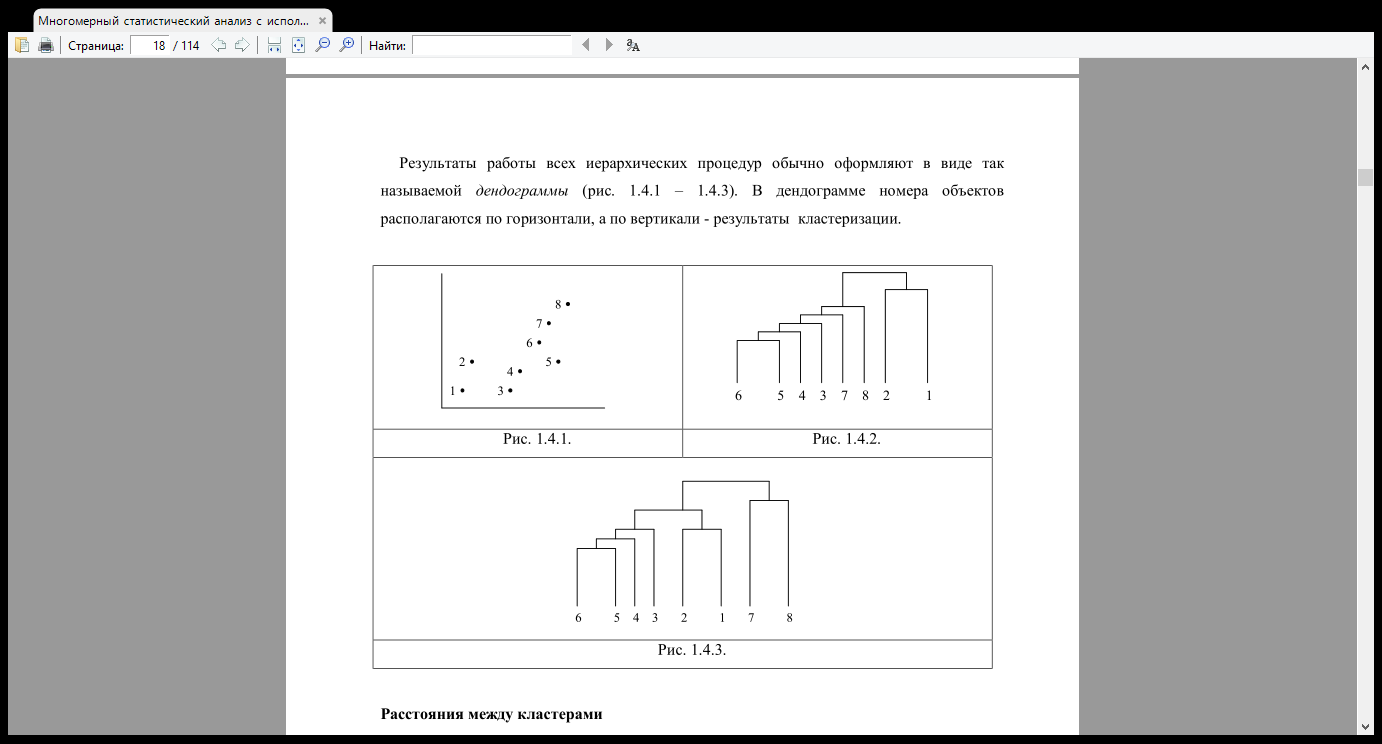


Рисунок 1.4.3

**Расстояния между кластерами**

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако когда связываются вместе несколько объектов, возникает вопрос, как следует определить расстояния между кластерами? Другими словами, необходимо правило объединения или связи для двух кластеров.

Здесь имеются различные возможности: например, вы можете связать два кластера вместе, когда любые два объекта в двух кластерах ближе друг к другу, чем соответствующее расстояние связи. Другими словами, вы используете правило ближайшего соседа для определения расстояния между кластерами; этот метод называется методом одиночной связи. Это правило строит волокнистые кластеры, т.е. кластеры, сцепленные вместе только отдельными элементами, случайно оказавшимися ближе остальных друг к другу. Как альтернативу вы можете использовать соседей в кластерах, которые находятся дальше всех остальных пар объектов друг от друга. Этот метод называется метод полной связи. Существует также множество других методов объединения кластеров, подобных тем, что были рассмотрены здесь, и модуль Кластерный анализ предлагает широкий выбор таких методов.

1. *Расстояние “Ближайшего соседа” (Одиночная связь).* Первый шаг 1.–7. совпадает с первым шагом алгоритма *Обобщённая агломеративная процедура*. Расстояние равно расстоянию между ближайшими объектами классов.

2. *Расстояние “Дальнего соседа” (Полная связь).* Расстояние равно расстоянию между самыми дальними объектами классов.

3. Невзвешенное попарное среднее. В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные рощи, однако он работает одинаково хорошо и в случаях протяженных (цепочного типа) кластеров.

4. *Взвешенное попарное среднее.* Метод идентичен методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому предлагаемый метод должен быть использован (скорее даже, чем предыдущий), когда предполагаются неравные размеры кластеров.

5. *Невзвешенный центроидный метод*. В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

6. *Взвешенный центроидный метод (медиана).* Тот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них). Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

7. *Метод Варда (Ward, 1963).* В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов (SS). Этот метод направлен на объединение близко расположенных кластеров.

1.3.2 Процедуры эталонного типа

Наряду с иерархическими методами классификации, существует многочисленная группа так называемых итеративных методов кластерного анализа (метод k-средних.). Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т.д.). Название метода было предложено Дж. Мак-Куином в 1967 г. В отличие от иерархических процедур, метод k-средних не требует вычисления и хранения матрицы расстояний или сходств между объектами. Алгоритм этого метода предполагает использование только исходных значений переменных. Для начала процедуры классификации должны быть заданы k выбранных объектов, которые будут служить эталонами, т.е. центрами кластеров. Считается, что алгоритмы эталонного типа удобные и быстродействующие. В этом случае важную роль играет выбор начальных условий, которые влияют на длительность процесса классификации и на его результаты. Метод k-средних удобен для обработки больших статистических совокупностей.

Математическое описание алгоритма метода k-средних (K-means)

Пусть имеется n наблюдений, каждое из которых характеризуется m признаками X1, X2, …, Xn. Эти наблюдения необходимо разбить на k кластеров. Для начала из n точек исследуемой совокупности отбираются случайным образом или задаются исследователем исходя из каких-либо априорных соображений k точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера. На первом шаге из оставшихся (n - k) объектов извлекается точка Xi с координатами (xi1, xi2, …, xim) и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, например, евклидово расстояние. Проверяемый объект присоединяется к тому центру (эталону), которому соответствует минимальное из расстояний. Эталон заменяется новым, пересчитанным с учетом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу. Если встречаются два или более минимальных расстояния, то i-ый объект присоединяют к центру с наименьшим порядковым номером. На следующем шаге выбираем точку Xi+1 и для нее повторяются все процедуры. Таким образом, через (n-k) шагов все точки (объекты) совокупности окажутся отнесенными к одному из k кластеров, но на этом процесс разбиения не заканчивается. Для того чтобы добиться устойчивости разбиения по тому же правилу, все точки X1, X2, …, Xn опять подсоединяются к полученным кластером, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется. Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить C1, C2, …, Ck. При этом каждая точка Xi (i = 1, 2, …, n) будет относиться к такому кластеру (классу) *l*, для которого расстояние минимально. Возможны две модификации метода *k*-средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, а вторая – лишь после того, как будет завершен просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется.