Элементы кластерного анализа

Учебное пособие

Благовещенск
2006

Федеральное агентство образования

АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет математики и информатики

Г. Н. Торопчина, Н. Н. Двоерядкина,
Г. П. Вохминцева

Элементы кластерного анализа

Учебное пособие

Благовещенск
2006

ББК 22. 161я73

Печатается по решению
редакционного - издательского совета
факультета математики и информатики
Амурского государственного
университета

Торопчина Г. Н., Двоерядкина Н. Н., Вохминцева Г. П.

Элементы кластерного анализа. Учебное пособие. Благовещенск: Амурский гос. ун-т, 2006.

Пособие предназначено для студентов второго курса. В нем рассматриваются теоретические сведения, решения типовых задач, задания для самостоятельной работы.

Рецензент: С. В. Ланкин, зав. Кафедрой общей физики БГПУ,
доктор физ.-мат. наук.

Научный редактор – д-р техн. наук проф. Г. В. Литовка.

Элементы кластерного анализа

§1. Введение в кластерный анализ

Классификация является основой умозрительной человеческой деятельности и фундаментальным процессом научной практики. В ходе исследований, развития науки и техники накоплено значительное количество материалов, которые необходимо систематизировать с целью выявления законов общественного развития, изучения эволюций и совершенствования технологий. Эта работа требует от исследователя детального изучения данных и их обобщения, в ходе которого отдельные факты складываются в закономерности, а закономерности в теории. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры. В настоящее время существует множество подходов к классификации объектов.

Среди них методы многомерного анализа - наиболее действенный количественный инструмент исследования социально - экономических процессов, описываемых большим числом характеристик. К ним относятся кластерный анализ, таксономия, распознавание образов, факторный анализ.

Кластерный анализ наиболее ярко отражает черты многомерного анализа в классификации, факторный анализ - в исследовании связи.

Термин кластерный анализ включает в себя набор различных алгоритмов классификации.

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. 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 позволяют решать следующие задачи:

1. Проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение такой задачи, как правило, приводит к углублению знаний о совокупности классифицируемых объектов;

2. Проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, т. е. поиск существующей структуры;

3. Построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

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 информации, делать их компактными и наглядными.

Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, характеризующих экономическое развитие. 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 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.

Поэтому необходимо сделать несколько предостережений общего характера.

1) Многие методы кластерного анализа - довольно простые эвристические процедуры, которые, как правило, не имеют достаточного статистического обоснования.

2) Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных. Это обычное явление в большинстве прикладных исследований.

3) Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, т. е. методы кластеризации могут приводить к порождению артефактов.

Исследования, использующие кластерный анализ, характеризуют следующие пять основных шагов: 1) отбор выборки для кластеризации; 2) определение множества признаков, по которым будут оцениваться объекты в выборке, и способа их стандартизации; 3) вычисление значений той или иной

меры сходства между объектами; 4) применение метода кластерного анализа для создания групп сходных объектов; 5) проверка достоверности результатов кластерного решения.

В кластерном анализе считается, что:

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

б) единицы измерения (масштаб) выбраны правильно.

§ 2. Задача кластерного анализа

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве *X,* разбить множество объектов *G* на *m* (*m* - целое) кластеров (подмножеств) $Q_1, Q_2, ..., Q_m$, так, чтобы каждый объект G_j 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.

Например, пусть *G* включает *n* стран, любая из которых характеризуется ВНП на душу населения (F_1), числом *M* автомашин на 1 тысячу человек (F_2), душевым потреблением электроэнергии (F_3), душевым потреблением стали (F_4) и т.д. Тогда X_1 (вектор измерений) представляет собой набор указанных характеристик для первой страны, X_2 - для второй, X_3 для третьей, и т.д. Задача заключается в том, чтобы разбить страны по уровню развития.

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 отклонений:

$$W = \sum_{j=1}^{n}(x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}(\sum_{j=1}^{n} x_j)^2$$

где x_j - представляет собой измерения j - го объекта.

Наиболее трудным в задаче классификации является определение меры однородности объектов.

Понятно, что объекты *i*-ый и *j*-ый попадали бы в один кластер, когда расстояние (отдаленность) между точками X_i и X_j было бы достаточно маленьким и попадали бы в разные кластеры, когда это расстояние было бы достаточно большим. Таким образом, попадание в один или разные кластеры объектов определяется понятием расстояния между X_i и X_j из E_p, где

E_p p - мерное евклидово пространство.

Неотрицательная функция $\rho(X_i, X_j)$ называется функцией расстояния (метрикой), если:

а) $\rho(X_i, X_j) \geq 0$, для всех X_i и X_j из E_p

б) $\rho(X_i, X_j) = 0$, тогда и только тогда, когда $X_i = X_j$

в) $\rho(X_i, X_j) = \rho(X_j, X_i)$

г) $\rho(X_i, X_j) \leq \rho(X_i, X_k) + \rho(X_{k,} X_j)$, где X_i; X_j и X_k - любые три вектора из E_p.

Значение $\rho(X_i, X_j)$ для X_i и X_j называется расстоянием между X_i и X_j и эквивалентно расстоянию между G_i и G_j соответственно выбранным характеристикам $(F_1, F_2, F_3, ..., F_p)$.

Наиболее часто употребляются следующие функции расстояний:

1. Евклидово расстояние $\rho_2(X_{i,}X_j) = \left[\sum_{k=1}^{p}(x_{ki} - x_{kj})^2\right]^{\frac{1}{2}}$.

2. l_1 - норма (манэттерское расстояниу) $\rho_1(X_i, X_j) = \left[\sum_{k=1}^{p}|x_{ki} - x_{кj}|\right]$.

3. Сюпремум – норма $\rho_{\infty}\left(X_i, X_j\right) = \sup\left\{\left|x_{ki} - x_{kj}\right|\right\}$ $k = 1, 2, \ldots, p.$

4. l_p - норма (степенное расстояие) $\rho_p\left(X_i, X_j\right) = \left[\sum_{k=1}^{p}\left|x_{ki} - x_{kj}\right|^p\right]^{\frac{1}{r}}$.

Евклидова метрика является наиболее популярной. Метрика l_1 наиболее легкая для вычислений. Сюпремум - норма легко считается и включает в себя процедуру упорядочения, а l_p - норма охватывает функции расстояний 1, 2,3,.

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

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

Тогда расстояние между парами векторов $\rho\left(X_i, X_j\right)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 0 & \ldots & \rho_n \\ \rho_{n1} & \rho_{n2} & \ldots & 0 \end{pmatrix}$$

Понятием, противоположным расстоянию, является понятие сходства между объектами G_l и G_j. Неотрицательная вещественная функция $S\left(X_i; X_j\right) = s_{ij}$ называется мерой сходства, если:

1) $0 \le S\left(X_i, X_j\right) < 1$ для $X_l \ne X_j$

2) $S\left(X_i, X_i\right) = 1$

3) $S\left(X_i, X_j\right) = S\left(X_j, X_i\right)$

Пары значений мер сходства можно объединить в матрицу сходства:

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

Величину s_{ij} называют коэффициентом сходства.

Естественной мерой сходства характеристик объектов во многих задачах является коэффициент корреляции между ними

$$r_{ij} = \frac{\sum_{h=1}^{N} (x_{hi} - m_i)(x_{hj} - m_j)}{\sigma_i \cdot \sigma_j}$$

где $m_i, m_j, \sigma_i, \sigma_j$ - соответственно средние и среднеквадратичные отклонения для характеристик *i* и *j*. Мерой различия между характеристиками может служить величина *1-r*.

§ 3. Данные

Кластерный анализ можно применять к интервальным данным, частотам, бинарными данным. Важно, чтобы переменные изменялись в сравнимых шкалах.

Неоднородность единиц измерения и вытекающая отсюда невозможность обоснованного выражения значений различных показателей в одном масштабе приводит к тому, что величина расстоянии между точками, отражающими положение объектов в пространстве их свойств, оказывается зависящей от произвольно избираемого масштаба. Чтобы устранить неоднородность измерения исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений к некоторой величине, отражающей определенные свойства данного показателя. Нормирование исходных данных для кластерного анализа иногда проводится посредством деления исходных величин на' среднеквадратичное отклонение соответствующих показателей. Другой способ сводится к вычислению, так называемого, стандартизованного вклада. Его еще называют *Z*-вкладом.

Z - вклад показывает, сколько стандартных отклонений отделяет данное

наблюдение от среднего значения:

$Z_i = \frac{x_i - \bar{x}}{\sigma_i}$, где x_i - значение данного наблюдения, $\bar{x}$ - среднее, σ_i - стандартное отклонение.

Среднее для *Z*-вкладов является нулевым и стандартное отклонение равно 1.

Стандартизация позволяет сравнивать наблюдения из различных распределений.

Заметим, что методы нормирования означают признание всех признаков равноценными с точки зрения выяснения сходства рассматриваемых объектов. Применительно к экономике признание равноценности различных показателей кажется оправданным отнюдь не всегда. Было бы желательным наряду с нормированием придать каждому из показателей вес, отражающий его значимость в ходе установления сходств и различий объектов.

В этой ситуации приходится прибегать к способу определения весов отдельных показателей - опросу экспертов.

Экспертные оценки дают известное основание для определения важности индикаторов, входящих в ту или иную группу показателей. Умножение нормированных значений показателей на коэффициент, соответствующий среднему баллу оценки, позволяет рассчитывать расстояния между точками с учетом неодинакового веса их признаков.

Довольно часто при решении подобных задач используют не один, а два расчета: первый, в котором все признаки считаются равнозначными, второй, где им придаются различные веса в соответствии со средними значениями экспертных оценок.

§ 4. Методы кластерного анализа

Существует много методов кластерного анализа. Основным понятием кластер – процедур является расстояние ρ_{st} между кластерами K_s и K_t .

Пусть имеется матрица расстояний $\{l_{ij}\}_{n\times n}$ между n объектами и некоторое их разбиение $\{K_1, K_2, \ldots K_p\}$ на p кластеров. Программой пакета «STATISTICA» предусмотрены следующие виды расстояний.

1. Метод полных связей (метод «дальнего соседа»).

Суть данного метода в том, что два объекта, принадлежащих одной и той же группе (кластеру), имеют коэффициент сходства, который меньше некоторого порогового значения l. В терминах евклидова расстояния это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения l. Таким образом, l определяет максимально допустимый диаметр подмножества, образующего кластер.

2. Метод максимального локального расстояния (метод «ближнего соседа»)

$$\rho_{st} = \min_{i\in K_s,\, j\in K_t} l_{ij}$$

Каждый объект рассматривается как одноточечный кластер. Объекты группируются 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 предыдущем методе для любых пороговых значений.

3. Метод Ворда.

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. Этот метод направлен на объединение близко расположенных кластеров.

4. Центроидный метод.

Расстояние между двумя кластерами определяется как евклидово расстояние между центрами (средними) этих кластеров: $\rho_{st} = l\left(\bar{x}_{K_s},\ \bar{x}_K\right)$,

где - $\bar{x}_{K_s}$ - среднее арифметическое векторных наблюдений x_i при $i \in K_s$.

Кластеризация идет поэтапно. На каждом из *n-1* шагов объединяют два кластера *G* и π, имеющие минимальное значение ρ_{ij}^2 Если n_1 много больше n_2 то центры объединения двух кластеров близки друг к другу и характеристики второго кластера при объединении кластеров практически игнорируются. Иногда этот метод называют еще методом взвешенных групп.

Названные методы относятся к группе иерархических (деревообразующих) агломеративных (объединительных) методов.

Число алгоритмов методов кластерного анализа слишком велико. Все их можно подразделить на иерархические и неиерархические.

Иерархические алгоритмы связаны с построением дендограмм и делятся на:

а) агломеративные, характеризуемые последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров;

б) дивизимные (делимые), в которых число кластеров возрастает, начиная с одного, в результате чего образуется последовательность расщепляющих групп.

Алгоритмы кластерного анализа имеют сегодня хорошую программ-

ную реализацию, которая позволяет решить задачи самой большой последовательности.

Иерархические агломеративные методы — многошаговые методы, работающие в такой последовательности: на нулевом шаге за разбиение принимается исходная совокупность *n* элементарных кластеров, матрица расстояний между которыми $\{\rho_{ij}\}_{n\times n} = \{l_{ij}\}_{n\times n}$; на каждом следующем шаге происходит объединение двух кластеров K_s и K_t, сформированных на предыдущем шаге, в один кластер $K_s \cup K_t$ (будем его обозначать $K_{s\oplus t}$, при этом размерность матрицы расстояний уменьшается, по сравнению с

размерностью матрицы предыдущего шага, на единицу. При использовании вышеназванных алгомеративных методов рассчитать расстояние $\rho_{i,s\oplus t}$ между кластерами $K_{s\oplus t}$ и K_i $(i \neq s,t)$ можно, используя соответствующую методу формулу расстояния между кластерами, однако менее трудоемки расчеты по формуле: $\rho_{i,s\oplus t} = \alpha\rho_{is} + \beta\rho_{it} + \gamma\rho_{st} + \delta\left|\rho_{is} - \rho_{it}\right|$, (1)

в которой значения коэффициентов $\alpha, \beta, \gamma, \delta$ зависят от используемого метода:

Метод	α	β	γ	δ	$\rho_{i,s\oplus t}$
Ближний сосед (Одинарной связи)	0,5	0,5	0	0,5	$\min(\rho_{is},\rho_{it}) = \min\limits_{p\in K_i, j\in K_{s\oplus t}} l_{pj}$ (2)
Дольний сосед (Полных связей)	0,5	0,5	0	0,5	$\max(\rho_{is},\rho_{it}) = \max\limits_{p\in K_i, j\in K_{s\oplus t}} l_{pj}$ (3)
Средней связи	$\frac{n_s}{n_s+n}$	$\frac{n_s}{n_s+n}$	0	0	$\frac{n_s\rho_{is}+n_t\rho_{it}}{n_s+n_t} = \frac{1}{n_i n_{s\oplus t}} \sum\limits_{p\in K_i}\sum\limits_{j\in K_{s\oplus t}} l_{pj}$ (4)
Центроидный	$\frac{n_s}{n_s+n}$	$\frac{n_s}{n_s+n}$	$-\alpha\beta$	0	$\frac{n_s\rho_{is}+n_t\rho_{it}}{n_s+n_t} - \frac{n_s n_t \rho_{st}}{(n_s+n_t)^2} = l\left(\bar{x}_{K_i}, \bar{x}_{K_{s\oplus t}}\right)$ (5)

В последнем столбце: слева приведена формула подсчета $\rho_{i,s\oplus t}$ вытекающая из (1), а справа — вытекающая из принятого в соответствующем методе определения расстояний между кластерами.

Наиболее известный метод представления матрицы расстояний или сходства основан на идее дендограммы или диаграммы дерева. Дендограмму можно определить как графическое изображение результатов процесса последовательной кластеризации, которая осуществляется в терминах матрицы расстояний. С помощью дендограммы можно графически или геометрически изобразить процедуру кластеризации при условии, что эта

процедура оперирует только с элементами матрицы расстояний или сходства.

Существует много способов построения дендограмм. В дендограмме объекты располагаются вертикально слева, результаты кластеризации - справа. Значения расстояний или сходства, отвечающие строению новых кластеров, изображаются по горизонтальной прямой поверх дендограмм.

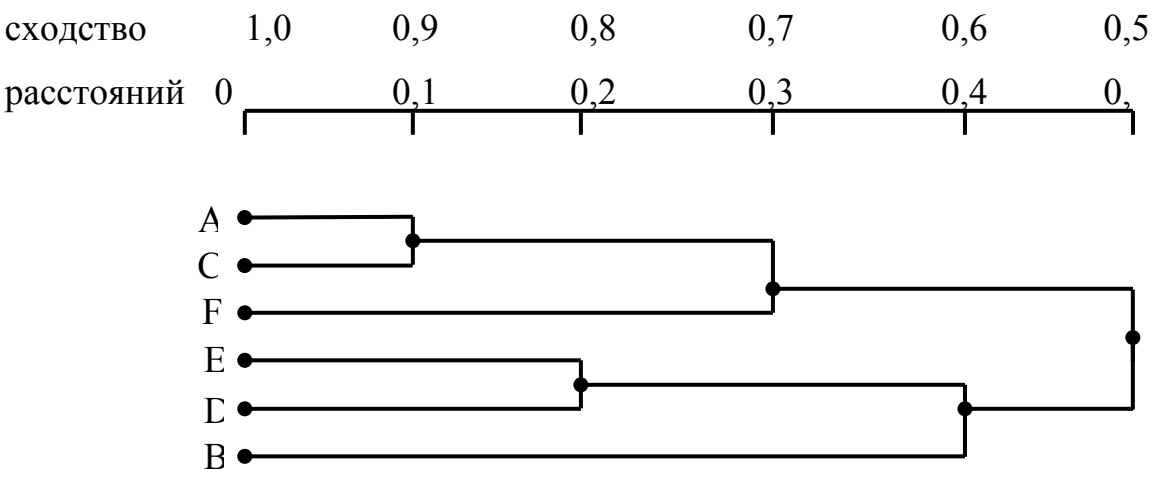

Рис. 1

На рисунке 1 показан один из примеров дендограммы. Рис. 1 соответствует случаю шести объектов *(n=6)* и *k* характеристик (признаков). Объекты *A* и *C* наиболее близки и поэтому объединяются в один кластер на уровне близости, равном 0,9. Объекты *D* и *E* объединяются при уровне 0.8. Теперь имеем 4 кластера:

$(A \oplus C), (F), (D \oplus E), (B)$.

Далее образуются кластеры $((A \oplus C) \oplus F)$ и $((E \oplus D), B)$, соответствующие уровню близости, равному 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне 0,5.

Вид дендограммы зависит от выбора меры сходства или расстояния между объектом и кластером и метода кластеризации.

Проиллюстрируем методы кластерного анализа на примере классификации пяти точек: (1,2), (4, 3), (-1, -1), (-1, 0), (-3, 3):

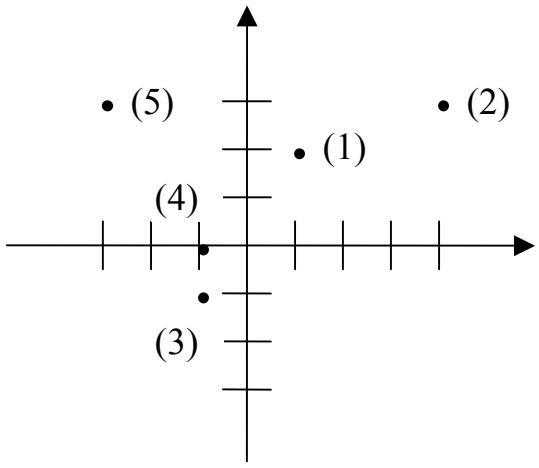

Рис. 2

За метрику расстояний примем квадратичное евклидово расстояние. Матрица расстояний:

	1	2	3	4	5
1	0	10	13	8	17
2	10	0	41	34	49
3	13	41	0	1	20
4	8	34	1	0	13
5	17	49	20	13	0

Начальное разбиение: 1, 2, 3, 4, 5. Минимальное расстояние $\rho_{3,4}=1,$ переходим к разбиению: 1, 2, $3\oplus 4$, 5.

Ближний сосед

	1	2	$3\oplus 4$	5
1	0	10	8	17
2	10	0	34	49
$3\oplus 4$	8	34	0	13
5	17	49	13	0

$\rightarrow$

	$1\oplus(3\oplus 4)$	2	5
$1\oplus(3\oplus 4)$	0	10	13
2	10	0	49
5	13	49	0

$\rightarrow$

	$(1\oplus(3\oplus 4))\oplus 2$	5
$(1\oplus(3\oplus 4))\oplus 2$	0	13
5	13	0

Последовательность разбиений: 1, 2, 3, 4, 5 $\rightarrow$ 1, 2, $3\oplus 4$, 5 $\rightarrow 1\oplus(3\oplus 4)$, 2, 5 $\rightarrow(1\oplus(3\oplus 4))\oplus 2$, 5 $\rightarrow 1\oplus 2\oplus 3\oplus 4\oplus 5$ представлена на рис. 3 а *дендрограммой* (сверху проставлены расстояния, при которых происходит переход к новому разбиению).

Поскольку этот метод объединяет кластеры, в которых расстояние между ближайшими элементами минимально по сравнению с другими кластерами, то два объекта попадают в один кластер, если существует соединяющая их цепочка ближайших друг к другу объектов («цепочечный эффект»). Поэтому метод «ближний сосед» называют методом «одиночной связи». Для устранения «цепочечного эффекта» можно, например, ввести ограничение на максимальное расстояние между объектами кластера: в первый кластер включить два наиболее близких объекта, затем в этот кластер включить объект, который имеет минимальное расстояние с одним из объектов кластера, а его расстояние до другого объекта кластера не больше числа l_0 и т. д.; формирование первого кластера продолжают до тех пор, пока нельзя будет найти объект, расстояние которого до любого объекта кластера, не превзойдет l_0 ; формирование 2-го и последующих кластеров осуществляется из оставшихся объектов аналогичным образом.

Дальний сосед

	1	2	$3\oplus4$	5
1	0	10	13	17
2	10	0	41	49
$3\oplus4$	13	41	0	20
5	17	49	20	0

$\rightarrow$

	$1\oplus2$	$3\oplus4$	5
$1\oplus2$	0	41	49
$3\oplus4$	41	0	20
5	49	20	0

$\rightarrow$

	$1\oplus2$	$(3\oplus4)\oplus5$
$1\oplus2$	0	49
$(3\oplus4)\oplus5$	49	0

Последовательность разбиений: $1, 2, 3, 4, 5 \rightarrow 1, 2, 3\oplus4, 5 \rightarrow 1\oplus2, 3\oplus4, 5 \rightarrow 1\oplus2, (3\oplus4)\oplus5 \rightarrow 1\oplus2\oplus3\oplus4\oplus5$ представлена на рис. 3б.

В этом методе объединяются кластеры, в которых минимально расстояние между самыми далекими друг от друга объектами. Это означает, что все остальные объекты в полученном после объединения кластере связаны друг с другом еще теснее, чем «соседи». Поэтому метод «дальнего соседа» называют методом полной связи.

Средняя связь

	1	2	$3\oplus4$	5
1	0	10	10.5	17
2	10	0	37.5	49
$3\oplus4$	10.5	37.5	0	16.5
5	17	49	16.5	0

$\rightarrow$

	$1\oplus2$	$3\oplus4$	5
$1\oplus2$	0	24	33
$3\oplus4$	24	0	16.5
5	33	16.5	0

$\rightarrow$

	$1\oplus2$	$(3\oplus4)\oplus5$
$1\oplus2$	0	27
$(3\oplus4)\oplus5$	27	0

В последней таблице $\rho_{1\oplus 2,(3\oplus 4)\oplus 5}=\frac{2}{3}\rho_{1\oplus 2,3\oplus 4}+\frac{1}{3}\rho_{1\oplus 2,5}=\frac{2}{3}\cdot 24+\frac{1}{3}\cdot 33=27$

Последовательность разбиений: $1,\ 2,\ 3,\ 4,\ 5\to 1,\ 2,\ 3\oplus 4,\ 5\to 1\oplus 2,$

$3\oplus 4,\ 5\to 1\oplus 2,(3\oplus 4)\oplus 5\to 1\oplus 2\oplus 3\oplus 4\oplus 5$ (см. рис. 3в).

Центроидный метод

	1	2	$3\oplus 4$	5
1	0	<u>10</u>	10.25	17
2	<u>10</u>	0	37.25	49
$3\oplus 4$	10.25	37.25	0	16.25
5	17	49	16.25	0

→

	$1\oplus 2$	$3\oplus 4$	5
$1\oplus 2$	0	21.25	30.5
$3\oplus 4$	21.25	0	<u>16.25</u>
5	30.5	<u>16.25</u>	0

→

	$1\oplus 2$	$(3\oplus 4)\oplus 5$
$1\oplus 2$ $(3\oplus 4)\oplus 5$	0 $\frac{373}{18}$	$\frac{373}{18}$ 0
	$1\oplus 2$	$(3\oplus 4)\oplus 5$

Дадим пояснение к расчётам некоторых расстояний. Расчёты проведём по тождественным формулам (5):

$$\rho_{1,3\oplus 4}=\begin{cases}\dfrac{\rho_{1,3}+\rho_{1,4}}{2}-\dfrac{\rho_{3,4}}{4}=\dfrac{13+8}{2}-\dfrac{1}{4}=10.25\\ l\left(\bar{x}_1,\bar{x}_{3\oplus 4}\right)=l\left((1,2),(-1,-0.5)\right)=(1+1)^2+(2+0.5)^2=10.25\end{cases}$$

$$\rho_{1\oplus 2,(3\oplus 4)\oplus 5}=\begin{cases}\dfrac{2\rho_{1\oplus 2,3\oplus 4}+\rho_{1\oplus 2,5}}{3}-\dfrac{2\rho_{3\oplus 4,5}}{9}=\dfrac{2\cdot 21.25+30.5}{3}-\dfrac{2\cdot 16.25}{9}=\dfrac{373}{18}\\ l\left(\bar{x}_{1\oplus 2},\bar{x}_{(3\oplus 4)\oplus 5}\right)=l\left(\left(\dfrac{5}{2},\dfrac{5}{2}\right),\left(-\dfrac{5}{3},\dfrac{2}{3}\right)\right)=\left(\dfrac{5}{2}+\dfrac{5}{3}\right)^2+\left(\dfrac{5}{2}-\dfrac{2}{3}\right)^2=\dfrac{373}{18}\end{cases}$$

Последовательность разбиений: $1,\ 2,\ 3,\ 4,\ 5\to 1,\ 2,\ 3\oplus 4,\ 5\to 1\oplus 2,\ 3\oplus 4,$ $5\to 1\oplus 2,(3\oplus 4)\oplus 5\to 1\oplus 2\oplus 3\oplus 4\oplus 5$ (см. рис. 3г).

Если число кластеров заранее известно, то классификацию заканчивают как только будет сформировано разбиение с этим числом кластеров. При неизвестном числе кластеров правило остановки связывают с понятием *порога* (уже использующимся ранее) – это некоторое расстояние l_0, определяемое условиями конкретной задачи. Например, если $l_0=\sum_{j>i}\frac{l_{ij}}{n(n-1)}$, в условиях

примера l_0=20.6, то метод «ближний сосед» объединит все пять объектов в один кластер, а остальные три метода дадут два кластера: $3 \oplus 4 \oplus 5$ и $1 \oplus 2$.

§ 5. Параллельные кластер-процедуры. Методы, связанные с функционалами качества разбиения

Иерархические методы используются обычно в задачах классификации небольшого числа объектов (порядка нескольких десятков), где больший интерес представляет не число кластеров, а анализ структуры множества этих объектов и наглядная интерпретация проведенного анализа в виде дендограммы. Если же число кластеров заранее задано или подлежит определению, то для классификации чаще всего используют *параллельные* кластер-процедуры — это итерационные алгоритмы, на каждом шаге которых *используется* одновременно (параллельно) все наблюдения. Так как эти алгоритмы на каждом шаге работают со всеми наблюдениями, то основной целью их конструирования является нахождение способов сокращения числа перебора вариантов (даже при числе наблюдений порядка нескольких десятков), что приводит зачастую лишь к приближенному, но не слишком трудоемкому решению задач. В параллельных кластер-процедурах реализуется обычно идея оптимизации разбиения в соответствии с некоторым функционалом качества.

Замечание. В рассмотренных выше иерархических методах реализованы схемы объединения объектов на основе эвристических соображений. Однако, например, для методов «ближнего соседа» и «дальнего соседа» можно указать функционалы качества *f(R)* разбиения *R*.

«Ближний сосед»: $f(R) = \max\limits_{p} \min\limits_{i,j \in K_p} l_{ij}$ (6)

«Дальний сосед»: $f(R) = \max\limits_{p} \max\limits_{i,j \in K_p} l_{ij}$ (7)

Наиболее распространенными являются при заданном числе k кластеров следующие функционалы качества разбиения:

- сумма внутрикластерных дисперсий

$$f(R)=\sum_{p=1}^{k}\sum_{i\in K_p} l^2\left(x_i,\bar{x}_{K_p}\right), \tag{8}$$

- сумма попарных внутрикластерных расстояний

$$f(R)=\sum_{p=1}^{k}\frac{1}{n_p}\sum_{i,j\in K_p} l^2\left(x_i,x_j\right), \tag{9}$$

а при неизвестном числе кластеров функционалы

$$f(R)=\alpha f_1(R)+\beta f_2(R)$$

или

$$f(R)=[f_1(R)]^{\alpha}[f_2(R)]^{\beta}$$

где $f_1(R)$ - некоторая не возрастающая функция числа классов, характеризующая средний внутриклассовый разброс наблюдений, $f_2(R)$ - некоторая неубывающая функция числа классов, характеризующая взаимную удаленность классов или меру «концентрации» наблюдений.

Схема работы алгоритмов, связанная с функционалами качества, такая: для некоторого начального разбиения R_0 вычисляют значение $f(R_0)$; затем каждую из точек x_i , поочередно перемещают во все кластеры и оставляют в том положении, которое соответствует наилучшему значению функционала качества. Работу заканчивают, когда перемещение точек не дает улучшения качества. Часто описанный алгоритм применяют несколько раз, начиная с разных начальных разбиений R_0 , и выбирают наилучший вариант разбиения.

§ 6. Число кластеров

Очень важным вопросом является проблема выбора необходимого числа кластеров. Иногда можно число кластеров выбирать априорно. Однако в общем случае это число определяется в процессе разбиения множества на кластеры.

Проводились исследования Фортьером и Соломоном, и было установлено, что число кластеров должно быть принято для достижения вероятности a того, что найдено наилучшее разбиение. Таким образом, оптимальное число разбиений является функцией заданной доли β наилучших или в некотором смысле допустимых разбиений во множестве всех возможных. Общее рассеяние будет тем больше, чем выше доля β допустимых разбиений. Фортьер и Соломон разработали таблицу, по которой можно найти число необходимых разбиений $S(\alpha,\beta)$ в зависимости от α и β (где α - вероятность того, что найдено наилучшее разбиение, β - доля наилучших разбиений в общем числе разбиений) Причем в качестве меры разнородности используется не мера рассеяния, а мера принадлежности, введенная Хользенгером и Харманом. Таблица значении $S(\alpha,\beta)$ приводится ниже.

Таблица значений $S(\alpha,\beta)$

$\beta \setminus \alpha$	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

Довольно часто критерием объединения (числа кластеров) становится изменение соответствующей функции. Например, суммы квадратов отклонений: $E_j = \sum_{i=1}^{n} r_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} r_{ij}\right)^2$

Процессу группировки должно соответствовать здесь последовательное минимальное возрастание значения критерия *E.* Наличие резкого скачка в

значении E можно интерпретировать как характеристику числа кластеров, объективно существующих в исследуемой совокупности.

Итак, второй способ определения наилучшего числа кластеров сводится к выявлению скачков, определяемых фазовым переходом от сильно связанного к слабосвязанному состоянию объектов.

§ 7. Последовательные кластер-процедуры

Метод *К* – средних

Иерархические и параллельные кластер-процедуры практически реализуемы лишь в задачах классификации не более нескольких десятков наблюдений. К решению задач с большим числом наблюдений применяют *последовательные* кластер-процедуры - это итерационные алгоритмы, на каждом шаге которых используется одно наблюдение (или небольшая часть исходных наблюдений) и результаты разбиения на предыдущем шаге. Идею этих процедур поясним на представленном в ППП «STASTICA» *методе К-средних* («*K* – Means Clusyering») с заранее заданным числом k классов.

На нулевом шаге за центры искомых k кластеров принимают случайно выбранные k наблюдений — точки $x_1, x_2, \ldots x_k$; каждому кластеру присваивают единичный вес. На первом шаге находят расстояния точки x_{k+1} до центров кластеров и точку x_{k+1} относят к кластеру, расстояние до которого минимально; рассчитывают новый центр тяжести (как взвешенное среднее по каждому показателю) этого кластера и вес кластера увеличивают на единицу; все остальные кластеры остаются неизмененными (с прежними центрами и весами). На втором шаге аналогичную процедуру выполняют для точки x_{k+2} и т. д. При достаточно большом числе n классифицируемых объектов или достаточно большом числе итерации пересчет центров тяжести практически не приводит к их изменению.

Если в какой-то точке не удается, прогнав все $x_{k+(n-1)}$ точек, достичь практически не изменяющихся центров тяжести, то либо используя получившееся

разбиение n точек на k кластеров в качестве начального применяют изложенную процедуру к точкам x_1, x_2 и т. д.; либо в качестве начального разбиения принимают различные комбинации k точек из исходных n точек и в качестве окончательного берут наиболее часто встречающееся финальное разбиение.

Реализация методов кластерного анализа с помощью пакета STATISTICA

Рассмотрим реализацию методов кластерного анализа с помощью компьютерного пакета STATISTICA. Данная программа позволяет осуществлять иерархический агломеративный метод (joining (tree clustering)), результатом которого является дендограмма, и последовательный итерационный метод (k-means clustering) с заранее заданным количеством кластеров.

Перед вызовом процедуры кластеризации необходимо стандартизировать данные, для того чтобы привести значения всех переменных к единому диапазону значений. Стандартизация осуществляется путем выбора процедуры Standardize (стандартизация) в пункте меню Data (переменные). Открывшееся диалоговое окно позволяет выбрать переменные (variables) для стандартизации и при необходимости придать вес (weight) некоторым переменным. Диалоговое окно стандартизации представлено на рисунке 1.

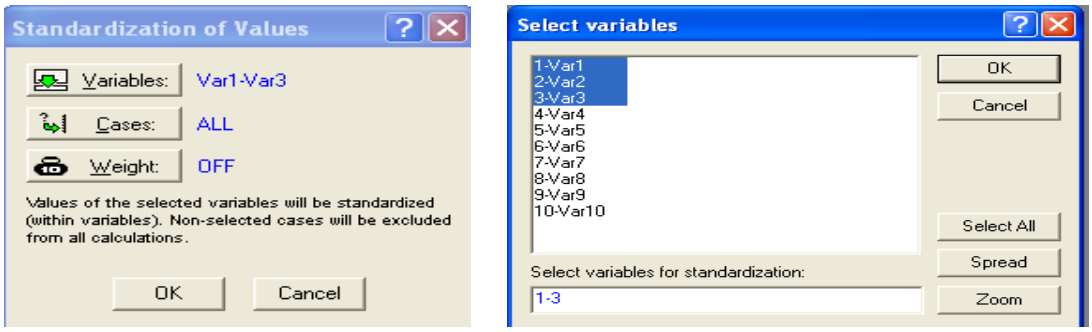

Рис.1 – Диалоговое окно для стандартизации выбранных переменных.

Вызов процедуры кластерного анализа осуществляется путем выбора пунктов меню Statistica / Multivariate Exploratory Techniques / Cluster Analysis, в

результате чего появляется диалоговое окно, позволяющее выбрать метод кластеризации (рис.2).

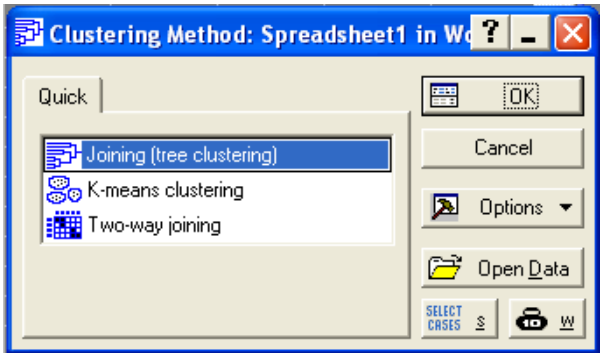

Рис.2 – Выбор методов кластеризации.

При реализации каждого из этих методов необходимо выбрать переменные, требующие группировки, а также указать тип кластеризации, т.к. в программе STATISTICA предусмотрена кластеризация как наблюдений (cases) так и переменных (variables). Рассмотрим каждый метод на примере.

Иерархический кластерный анализ

В иерархическом методе каждое наблюдение образовывает сначала свой отдельный кластер. На первом шаге два соседних кластера объединяются в один; этот процесс может продолжаться до тех пор, пока не останутся только два кластера.

Рассмотрим построение дендограммы на примере из области кадровой политики некоего предприятия. 18 претендентов прошли 10 различных тестов в кадровом отделе предприятия.

Таблица 1 - Список тестов.

Номер теста	**обозначение**	**Предмет теста**
1	t1	Память на числа
2	t2	Математические задачи
3	t3	Находчивость при прямом диалоге
4	t4	Тест на составление алгоритмов
5	t5	Уверенность во время выступления
6	t6	Командный дух
7	t7	Находчивость
8	t8	Сотрудничество

9	t9	Признание в коллективе
10	t10	Сила убеждения

Максимальная оценка, которую можно было получить на каждом из тестов, составляет 10 балов. Результаты теста находятся в таблице 2 в переменных t1-t10. Каждое наблюдение является характеристикой тестируемых кандидатов.

Таблица 2 - Результаты тестов 18 претендентов.

Имя	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
Andrew	10	9	3	10	4	5	1	7	6	5
Mark	6	5	8	6	9	6	2	6	5	9
Carol	7	4	5	6	6	7	5	5	4	6
Daniel	8	8	5	6	6	4	6	7	7	5
Victor	5	4	8	3	9	2	9	9	8	9
Lion	6	4	4	5	6	8	7	4	2	6
Erica	9	10	7	8	6	9	3	6	5	7
Martin	4	3	7	3	6	5	6	10	9	6
Bunny	8	5	4	9	5	8	4	6	7	7
Lola	8	6	4	6	5	7	8	6	8	5
Steve	7	7	6	7	7	9	9	3	4	7
Timothy	5	5	9	5	10	5	10	6	5	9
Susan	4	5	10	4	10	4	5	8	9	10
Denis	10	10	6	9	5	8	4	8	10	5
Scott	6	5	4	6	4	4	6	5	6	4
Gari	6	7	3	7	2	3	8	4	3	3
Laurie	7	8	5	7	3	6	7	4	4	4
Jeremy	8	9	6	8	5	10	7	10	8	5

С использованием результатов теста соответствия, мы хотим провести кластерный анализ, целью которого является обнаружение групп кандидатов, близких по своим качествам. С этой целью:

1. Создайте файл, содержащий информацию о тестируемых кандидатах. Так как все переменные в этом примере имеют одинаковые пределы изменения значений, стандартизация переменных является излишней.

2. Выберите в меню Statistica (Статистика) Multivariate Exploratory Techniques (Многомерные методы исследования) Cluster Analysis (Кластерный анализ) Joining(tree clustering) (Иерархический кластерный анализ).

3. В диалоговом окне Joining(tree clustering) (Иерархический кластерный анализ) переменные t1-t10 выберите в качестве тестируемых переменных, а наблюдения (cases) - имена кандидатов используйте для маркировки. По умолчанию в программе STATISTICA устанавливается Euclidean distances (евклидово расстояние) и Single Linkage (метод полных связей).

Диалоговое окно будет выглядеть как на рисунке 3.

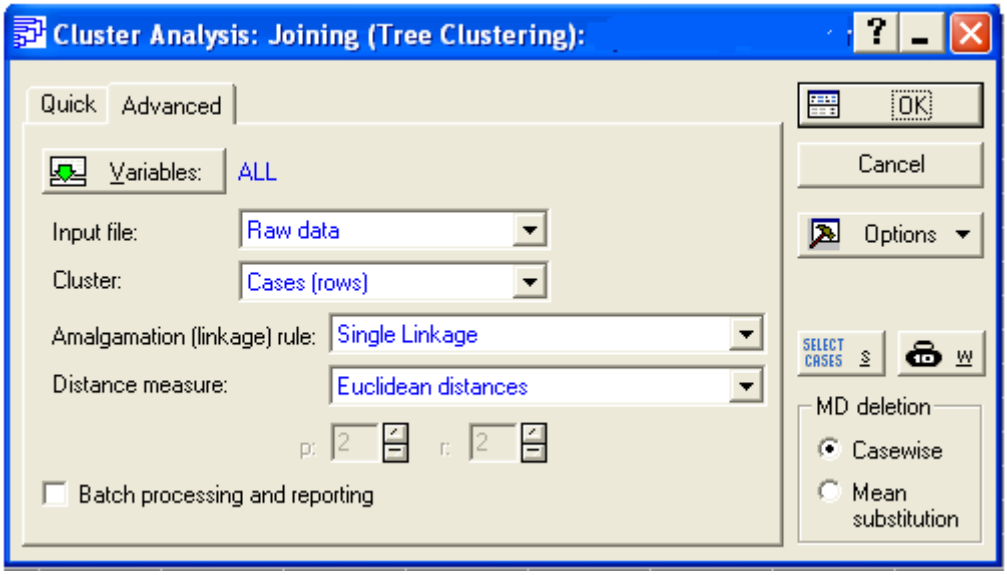

Рис.3 – Диалоговое окно иерархического кластерного анализа.

Кроме того, в данном диалоговом окне можно выбрать различные расстояния между объектами (Distance measure) и меры связи кластеров (Amalgamation (linkage) rule) .

Расстояния:

- Euclidean distances – евклидово расстояние;
- Squared euclidean distances – квадрат евклидова расстояния;
- City-block (Manhattan) distances – расстояние городских кварталов (манхэттенское расстояние);
- Chebychev distance metric – расстояние Чебышева;

- Power distance – степенное расстояние;
- Percent disagreement – процент несогласия;
- 1-Pearson r – (1- коэффициент корреляции Пирсона).

Методы связи кластеров:

- Single Linkage – одиночная связь (метод ближайшего соседа);
- Complete Linkage – полная связь (метод наиболее удаленных соседей);
- Unweighted pair-group average – невзвешенное попарное среднее;
- Weighted pair-group average – взвешенное попарное среднее;
- Unweighted pair-group centroid – невзвешенный центроидный метод;
- Weighted pair-group centroid (median) – взвешенный центроидный метод (медиана);
- Ward's method – метод Варда.

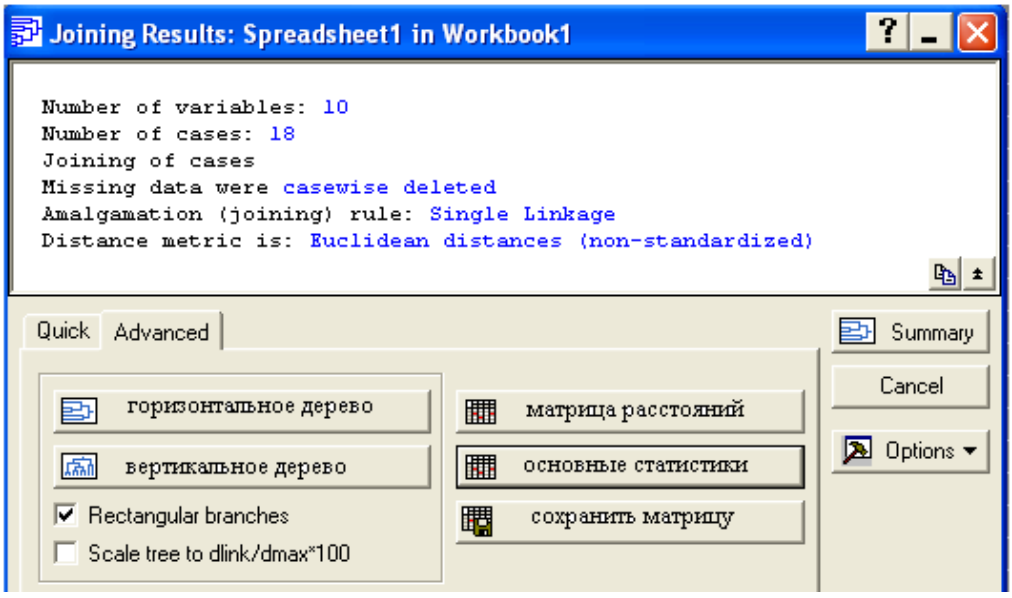

Рис. 4 – Диалоговое окно результатов иерархического кластерного анализа

Основные результаты иерархического кластерного анализа представлены в форме матрицы расстояний и горизонтального или вертикального дерева (дендограммы), которые вызываются путем нажатия соответствующих кнопок на диалоговом окне результатов кластерного анализа (рис.4).

Таблица 3 – Первоначальная матрица евклидовых расстояний.

	Andrew	Mark	Carol	Daniel	Victor	Lion	Erica	Martin	Bunny	Lola	Steve	Timothy	Susan
Andrew	0	11	9	7	16	12	7	14	7	9	12	15	1
Mark	11	0	6	8	10	9	8	9	8	10	9	8	
Carol	9	6	0	6	11	4	8	9	5	6	6	9	1
Daniel	7	8	6	0	9	9	7	8	7	5	8	9	1
Victor	16	10	11	9	0	12	14	6	12	10	12	6	
Lion	12	9	4	9	12	0	10	11	8	7	5	9	1
Erica	7	8	8	7	14	10	0	13	7	9	8	12	1
Martin	14	9	9	8	6	11	13	0	10	8	12	9	
Bunny	7	8	5	7	12	8	7	10	0	6	8	11	1
Lola	9	10	6	5	10	7	9	8	6	0	7	10	1
Steve	12	9	6	8	12	5	8	12	8	7	0	8	1
Timothy	15	8	9	9	6	9	12	9	11	10	8	0	
Susan	15	7	11	10	6	13	12	7	12	11	12	7	
Denis	7	11	10	7	14	13	6	12	7	8	11	14	1
Scott	9	9	5	5	10	7	10	8	7	5	8	10	1
Gari	10	13	8	8	14	8	12	13	10	8	9	13	1
Laurie	9	11	6	6	13	7	8	11	7	6	6	11	1
Jeremy	10	11	9	7	13	11	7	10	7	7	9	12	1

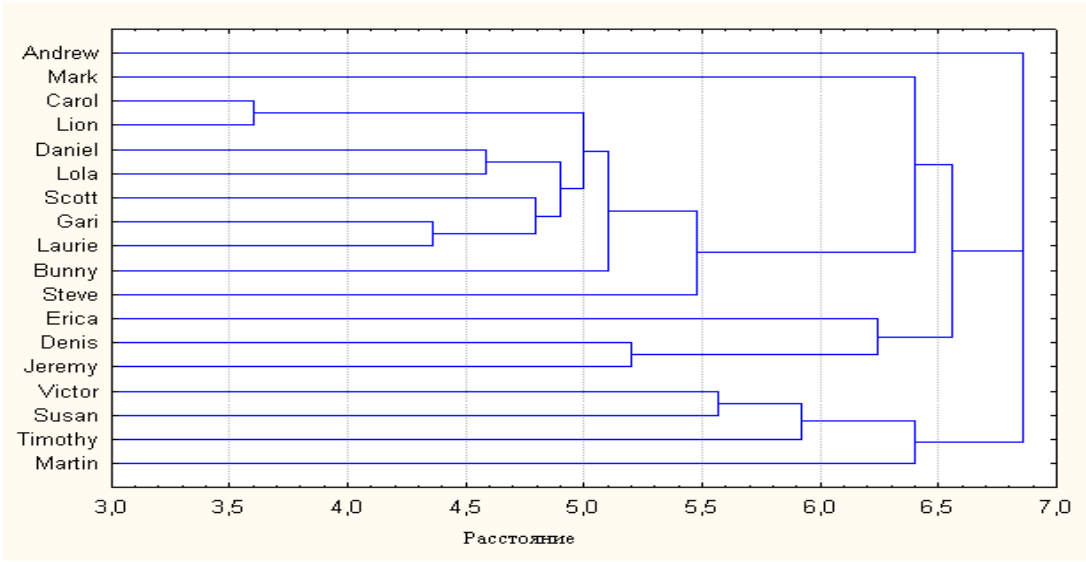

Рис.5 – Горизонтальная дендограмма.

Анализируя горизонтальную дендограмму можно предположить, что в один кластер входят четыре человека (Victor, Susan, Timothy, Martin), в другой кластер – три человека (Jeremy, Denis, Erica) , в третий кластер – десять

человек (Mark, Carol, Daniel, Lion, Bunny, Lola, Steve, Scott, Gari, Laurie) и в четвёртый кластер – один человек (Andrew). Неясно ещё, что означают эти четыре кластера, то есть о чём говорят результаты 10 тестов, соответственно относящиеся к этим кластерам.

Разобраться в значении кластеров помогают кластерные профили; они представляют собой средние значения переменных, которые включены в анализ, распределённые по кластерной принадлежности.

Определим средние значения всех переменных по каждому из выделенных кластеров.

Таблица 4 – Средние значения наблюдений по тестам в кластерах.

Предмет теста	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Память на числа	4,5	9	6,9	10
Математические задачи	4,3	9,7	5,9	9
Находчивость при прямом диалоге	8,5	6,3	4,8	3
Тест на составление алгоритмов	3,8	8,3	6,5	10
Уверенность во время выступления	8,8	5,3	5,3	4
Командный дух	4	9	6,2	5
Находчивость	7,5	4,7	6,2	1
Сотрудничество	8,3	8	5	7
Признание в коллективе	7,8	7,7	5	6
Сила убеждения	8,5	5,7	5,6	5

В первый кластер вошли испытуемые, которые уверенно себя чувствуют во время выступления, но имеют слабые показатели в математических тестах. Во второй кластер включены те, кто имеет очень хорошие показатели по математическим тестам (память на числа, математические задачи, тест на составление алгоритмов), но со средними оценками в социальной компетентности и уверенности при выступлениях. Тестируемые, входящие в третий кластер имеют средние показатели во всех тестах. В четвёртом кластере,

собраны люди с высоким уровнем математических знаний, но со слабыми результатами в тестах на социальную компетентность и на силу убеждения.

Недостатком иерархического кластерного анализа является сложность, а зачастую невозможность интерпретации результатов полученной дендограммы. Поэтому иерархический анализ удобно использовать лишь при небольшом количестве наблюдений (не более 10). При большем количестве наблюдений иерархический анализ является разведочным методом для последовательного итерационного анализа. Он позволяет определить примерное количество кластеров разбиения.

Кластерный анализ при большом количестве наблюдений (Кластерный анализ методом к-средних)

Иерархические методы объединения, хотя и точны, но трудоёмки: на каждом шаге необходимо выстраивать дистанционную матрицу для всех текущих кластеров. Расчётное время растёт пропорционально третьей степени количества наблюдений, что при наличии нескольких тысяч наблюдений может утомить и серьёзные вычислительные машины.

Поэтому при наличии большого количества наблюдений применяют другие методы. Недостаток этих методов заключается в том, что здесь необходимо заранее задавать количество кластеров, а не так как в иерархическом анализе, получить это в качестве результата. Эту проблему можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров. Если количество кластеров указать предварительно, то появляется следующая проблема: определение начальных значений центров кластеров. Их также можно взять из предварительно проведённого иерархического анализа, в котором для каждого наблюдения рассчитывают средние значения переменных, использовавшихся при анализе, а потом в определённой форме сохраняют их в некотором файле. Этот файл может быть, затем прочитан методом, который применяется для обработки больших

количеств наблюдений. Если нет желания проходить весь этот длинный путь, то можно воспользоваться методом, предлагаемым для данного наблюдения программой STATISTICA.

Если количество кластеров k, которое необходимо получить в результате объединения, задано заранее, то первые k наблюдений, содержащихся в файле, используются как первые кластеры. На последующих шагах кластерный центр заменяется наблюдением, если наименьшее расстояние от него до кластерного центра больше расстояния между двумя ближайшими кластерами. По этому правилу заменяется тот кластерный центр, который находится ближе всего к данному наблюдению. Таким образом получается новый набор исходных кластерных центров. Для завершения шага процедуры рассчитывается новое положение центров кластеров, а наблюдения перераспределяются между кластерами с изменённым центрами. Этот итерационный процесс продолжается до тех пор, пока кластерные центры не перестанут изменять свое положение или пока не будет достигнуто максимальное число итераций.

В качестве примера расчёта по этому алгоритму, рассмотрим выборку из результатов тестирования 18 кандидатов некоторого предприятия, представленную выше.

1. Откройте файл, содержащий данные тестирования.

2. Выберите в меню Statistica (Статистика) Multivariate Exploratory Techniques (Многомерные методы исследования) Cluster Analysis (Кластерный анализ) k-means clustering (Кластерный анализ методом к - средних).

3. Откроется диалоговое окно K-Means Cluster Analysis (Кластерный анализ методом к - средних).

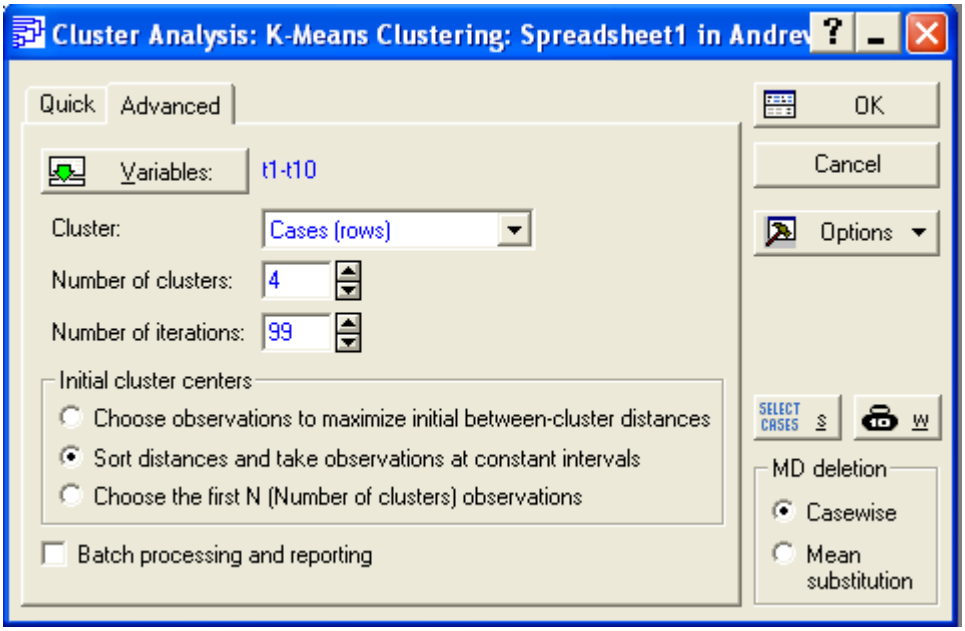

Рис.6 - Диалоговое окно K-Means Cluster Analysis (Кластерный анализ методом к-средних).

Переменные от t1 до t10 выберите в качестве тестируемых переменных. В поле Cluster укажите, что кластеризация проводится по наблюдениям (cases). Теперь необходимо указать количество кластеров (Number of cluster) и количество итераций (Number of iteration).

Укажите число итераций равное 99; установленное по умолчанию количество итераций равное 10, иногда оказывается недостаточным.

Для определения количества кластеров k можно провести несколько опытных, пробных расчётов с различным количеством кластеров и после этого определиться с подходящим вариантом решения. Но наиболее подходящим вариантом определения k является предварительное проведение иерархического кластерного анализа для произвольно выбранных наблюдений и получившееся количество кластеров принять за оптимальное.

Мы остановимся на четырёх кластерах, т.к. именно такое количество удалось выделить в предыдущем примере. Введем значение 4 в поле Number of Clusters (Количество кластеров).

Щёлкните на ОК, чтобы начать расчёт.

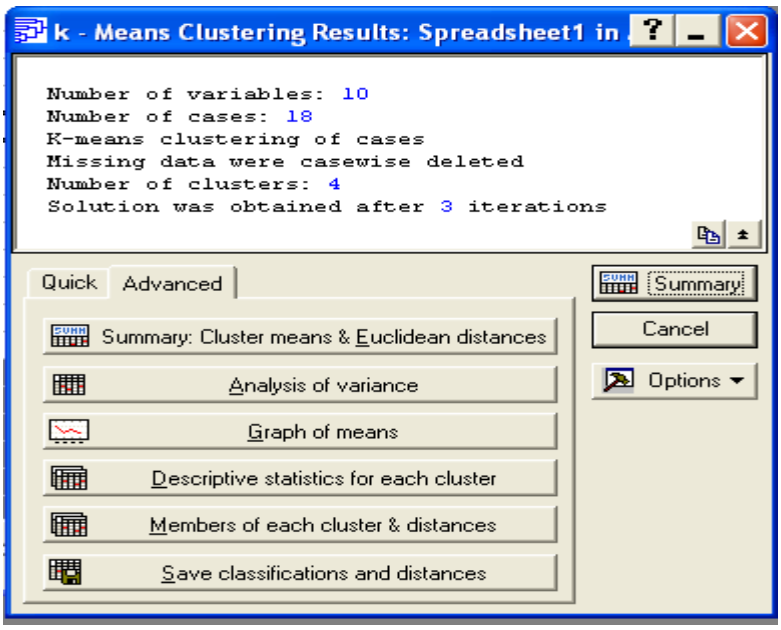

Рис.7 – Окно просмотра результатов кластерного анализа методом к-средних.

Сначала приводятся первичные кластерные центры (cluster means) и евклидовы расстояния между кластерами (Euclidean distances).

Таблица 5 – Средние значения переменных в каждом кластере.

переменная	клас тер1	клас тер2	клас тер3	клас тер4
Память на числа	4,8	8,0	6,5	9,7
Математические задачи	4,4	7,0	5,8	9,7
Находчивость при прямом диалоге	8,4	4,8	4,5	5,3
Тест на составление алгоритмов	4,2	7,3	6,3	9,0
Уверенность во время выступления	8,8	5,3	4,7	5,0
Командный дух	4,4	7,3	6,2	7,3
Находчивость	6,4	6,3	7,0	2,7
Сотрудничество	7,8	7,3	4,2	7,0
Признание в коллективе	7,2	7,5	3,8	7,0
Сила убеждения	8,6	5,5	5,0	5,7

Таблица 6 – Матрица евклидовых расстояний между кластерами.

	кластер1	кластер2	кластер3	кластер4
кластер1	0	7,04	8,28	13,01
кластер2	2,65	0	2,98	2,65
кластер3	2,88	1,73	0	7,13
кластер4	3,61	1,63	2,67	0

Далее выводятся показатели, позволяющие анализировать каждый кластер: анализ разброса значений (Analysis of variance), графическое представление кластерных центров (Graph of mean), основные статистики (среднее, стандартное отклонение и дисперсия) для каждого кластера (Descriptive statistics for each cluster).

В заключение можно просмотреть и сохранить наблюдения, относящиеся к каждому из кластеров и расстояния между наблюдениями. Полученная классификация и расстояния сохраняются в отдельном файле, в который для удобства пользователя и возможно для дальнейшей работы можно добавить необходимые переменные из исходного рабочего документа, путем выделения соответствующих строк в открывающемся диалоговом окне.

Итак, кластер 1 составляют: Mark, Victor, Martin, Timothy, Susan; кластер 2: Daniel, Bunny, Lola, Jeremy; кластер 3: Carol, Lion, Steve, Scott, Gari, Laurie; кластер 4: Andrew, Erica, Denis.

Кластерный анализ методом к-средних дополняет и уточняет картину, полученную с помощью иерархического кластерного анализа. Однако, конфигурация кластеров не поддается представлению в графическом виде.

§ 11. Задачи для самостоятельной работы

Задание 1. Используя не менее двух методов кластер – процедур провести классификацию и построить дендограммы для данных точек.

Точки / Варианты	A	B	C	D	E
1	(3; 4)	(-3; 8)	(-2;-6)	(5; 7)	(6; 0)
2	(8; 4)	(-3; 9)	(2;-6)	(5; -7)	(5; 0)
3	(-4; 4)	(-3; 0)	(5;-6)	(3; -7)	(7; 0)
4	(-5; 4)	(0; 8)	(-9;-6)	(-2; 7)	(6; 5)
5	(6; 4)	(-3; -1)	(4;-6)	(7; 7)	(6; 1)
6	(2; 4)	(-3; 5)	(-2;7)	(5; 3)	(6; 3)
7	(5; 4)	(-3; 0)	(7;-6)	(6; 7)	(9; 0)
8	(1; 4)	(-3; 2)	(-2;-7)	(4; 7)	(6; 3)
9	(5; 4)	(-3; 6)	(-2; 9)	(1; 7)	(0; 2)
10	(3; -5)	(3; -8)	(4;-6)	(-3; 7)	(1; 0)
11	(0; 4)	(-3; 0)	(4;-6)	(-1; 7)	(6; 6)
12	(0; -4)	(-3; 7)	(5;-6)	(-1; 7)	(5; 1)
13	(-3; 2)	(-6; 8)	(7;-6)	(0; 7)	(8; 1)
14	(3; -4)	(-3; 4)	(9;-6)	(5; 0)	(6; 4)
15	(-7; 4)	(-3; 5)	(-8;-6)	(0; 7)	(9; 1)
16	(8; 4)	(-3; 0)	(5;-6)	(5; 2)	(8; 2)
17	(0; 9)	(-3; 4)	(-5;-1)	(5; 0)	(3; 1)
18	(7; 0)	(-2; -6)	(-2;-6)	(5; 7)	(5; 4)
19	(5; -4)	(-3; 7)	(-4;-6)	(8; 6)	(2; -7)
20	(5; -3)	(6; -5)	(0;-6)	(1; 7)	(5; 0)
21	(2; 4)	(-3; 0)	(7;-6)	(4; 8)	(1; 3)
22	(4; 4)	(-6; 8)	(-2; 0)	(5; 7)	(6; 3)
23	(6; 4)	(-3; -2)	(-2; 0)	(5; 9)	(5; 2)
24	(-2; 4)	(-3; 8)	(-1;-6)	(5; 2)	(5; 3)
25	(3; -4)	(-3; 0)	(-2;0)	(5; 5)	(6; 2)

Задача 2. Исследуются магазины города, каждый из которых характеризуется рядом признаков, представленных в таблице. Разбить магазины на группы по уровню развития.

Таблица 7

Название	число клиентов сутки	кол-во касс	средняя цена на товар	Время обслужи-вания в сутки	число флиало в	тип	близ ость к цент ру
Проддвижени е	2000	3	350	24	1	1	
Любимый	1800	2	56	12	4	2	
Универмаг	3500	100	210	8	1	3	
Россия	2500	50	120	10	5	1	
Товары для дома	2800	50	250	8	1	2	
Фауст	5000	3	65	24	25	1	
S-стиль	200	1	3500	7	1	2	
Бенетон	350	1	2000	7	1	2	
Гастроном	1000	1	300	12	1	1	
Амурская ярмарка	3800	390	850	8	1	3	

$$\text{тип магазина} = \begin{cases} 1 - \textit{продуктовый} \\ 2 - \textit{непродуктовый} \\ 3 - \textit{универсальный} \end{cases}$$

Задача 3. Разбить представленные в таблице 8 страны по уровню экономического развития, если известны доли импорта и экспорта данных стран, общее количество населения, а также процент безработного населения каждой страны.

Таблица 8

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения

Россия	50	80	146	20
Украина	20	50	120	35
Белоруссия	30	25	100	15
Польша	10	60	95	10
Бельгия	60	20	70	15

Задача 4. Исследуются автомобили различных марок с целью разбить их по группам в зависимости от представленных в таблице 9 характеристик.

Таблица 9

Марка авто	Время разгона (с)	Максимальная скорость (км/ч)	Средняя цена (тыс.руб)	Престиж-ность (0-нет,1-да)	Число моделей (шт.)	Гарантийный срок (в годах)
Мазда	7	170	170	0	20	4
Тойота	8	180	120	1	28	5
Сузуки	13	160	130	0	15	3
Мицубиси	9	190	150	0	35	7
Форд	4	245	600	1	18	10
Рено	7	220	500	1	22	15
Джип	4	200	350	1	17	12
ВАЗ	20	140	60	0	31	2

Задача 5. Объединить 8 Амурских фирм, занимающихся производством и установкой окон в несколько схожих совокупностей. Характеристики фирм представлены в таблице 10.

Таблица 10

фирма	возраст фирмы	кол-во сотруд-ников	количество установленных окон за месяц	средняя цена окна (тыс.руб)	Количество филиалов по области	Скидки (%)	популяр-ность (0-нет, 1д -а)
Home master	11	200	30	17	20	25	1
Уют	3	14	12	13,2	2	33	1
РосОкна	3	50	8	14,8	5	30	1

Ремикс	5	80	14	15	10	10	1
Ванда	8	35	17	14,2	12	15	1
Ремстрой	2	10	5	11,5	1	5	0
СМУ-17	7	20	3	13,5	1	0	0
Оникс	3	20	1	12,9	1	0	0

Задача 6. При исследовании состояния производственной дисциплины в некоторой организации 15 работников, которые явились экспертами, оценили уровень значимости каждого из представленных в таблице 5 факторов, характеризующих дисциплину, по 10-бальной шкале (10-значим, 1-незначим). По оценкам экспертов сократить количество факторов, путем объединения родственных фактов в совокупности. Разбить экспертов в зависимости от их оценок.

Таблица 11

№ эксперта	организация труда	обеспечение сырьем	оплата труда	нормирование труда	условия труда	условия охраны труда	стабильность кадров	техническое обслуживание	качество планирования	социально-психологический климат	забота о бытовых нуждах работников
1	7	3	8	6	10	7	3	6	8	9	10
2	6	7	9	2	7	8	5	5	3	10	8
3	2	1	6	8	8	5	4	3	4	6	9
4	10	2	10	9	5	2	1	4	6	4	7
5	8	6	5	10	9	6	6	9	10	5	6
6	6	4	8	4	6	10	1	8	1	8	5
7	3	8	7	5	4	1	7	5	2	9	3
8	1	10	8	7	10	9	2	7	6	7	7
9	9	5	10	10	6	5	1	10	8	4	10
10	7	1	4	0	3	9	4	4	7	10	5
11	4	6	7	3	5	10	8	9	9	3	2
12	10	7	5	6	9	6	3	8	10	1	1
13	5	4	10	5	1	4	2	6	1	5	1

14	8	2	6	7	8	8	1	3	1	1	6
15	1	7	3	3	10	7	10	7	4	7	8

Задание 7. Провести небольшое исследование с применением кластерного анализа в получаемой специальности, для чего:

1. Выбрать объект исследования;
2. Провести интервью с целью получения перечня высказываний или свойств для дальнейшей оценки по интервальной или порядковой шкале;
3. Провести опрос (анкетирование);
4. Исследовать полученную информацию методами кластерного анализа;
5. Описать профили полученных кластеров;
6. Выработать рекомендации по исследуемому объекту.

Литература

1. Алексеев А. А. Маркетинговые исследования рынка услуг. Учеб. Пособие. − СПб.: СПбУЭФ, - 1998.
2. Алексеев А. А. «Методика сегментирования потребителей», // «Маркетинг и маркетинговые исследования в России», № 1,- 1999.
3. Голубков Е. П. «Сегментация и позиционирование».// «Маркетинг в России и за рубежом», № 4, - 2001.
4. Кастерин А. Г. Практика сегментирования рынка. − СПб.: Питер, 2002. – 288с.
5. Мотышина М. С. Методы и модели маркетинговых исследований: Учеб. пособие. − СПб.: СПбУЭФ, - 1996.
6. Попов Е. В. Теория маркетингового исследования. Екатеринбург.: УГУТУ.-1998. – 200с.
7. . Попов Е. В. «Сегментация рынка», // «Маркетинг в России и за рубежом», № 2, - 1999.
8. Пиотровский А. Л. Практический маркетинг. // Кластерный анализ как инструмент подготовки эффективных маркетинговых решений. №5,-

2001.− 2-5с.
9. htpp://www.segmentation.ru
10. htpp://www.iteam.ru

Оглавление

§1. Введение в кластерный анализ 3
§ 2. Задача кластерного анализа 6
§ 3. Данные 9
§ 4. Методы кластерного анализа 10
§ 5. Параллельные кластер-процедуры. Методы, связанные с функционалами качества разбиения 18
§ 6. Число кластеров 19
§ 7. Последовательные кластер-процедуры Метод K – средних 21
§ 8. Задачи для самостоятельной работы 33
Литература 38