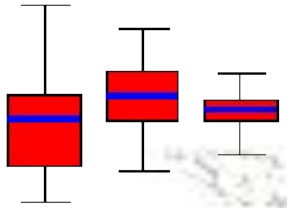
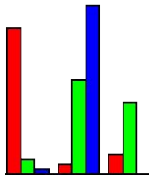
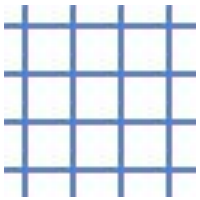
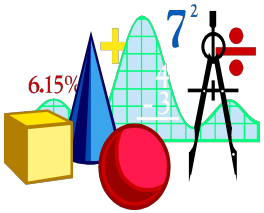




Кластерный анализ

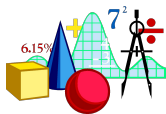


Стат. методы в психологии (Радчикова Н.П.)



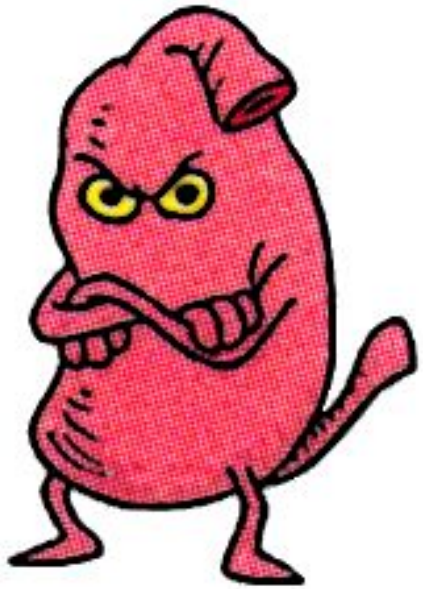
Цели

Что такое кластерный анализ и для чего он может понадобиться?



Кластерный анализ

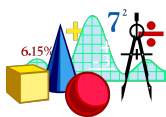
Если долго пытать данные, то они в конце концов сознаются…



Кластерный анализ

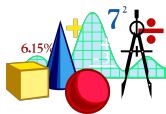
это общее название множества вычислительных процедур, используемых при создании классификации.

Главная цель кластерного анализа – нахождение групп схожих объектов в выборке данных. Эти группы удобно называть кластерами.



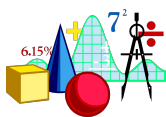
Кластерный анализ

Кластерный анализ – это метод, который позволяет разделить объекты СРАЗУ по нескольким характеристикам



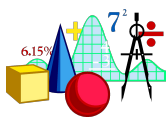
Кластерный анализ

Не существует общепринятого определения термина «кластер», однако считается, что кластеры обладают некоторыми свойствами, наиболее важными из которых являются плотность, дисперсия, размеры, форма и отделимость.



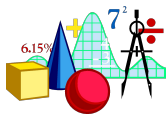
Свойства кластеров

Плотность – это свойство, которое позволяет определить кластер как скопление точек в пространстве данных, относительно плотное по сравнению с другими областями пространства, содержащими либо мало точек, либо не содержащими их вовсе.



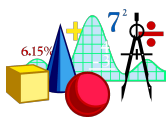
Свойства кластеров

Дисперсия характеризует степень рассеяния точек в пространстве относительно центра кластера, т.е. насколько близко друг к другу расположены точки кластера.



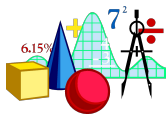
Свойства кластеров

Размеры тесно связано с дисперсией; если кластер можно идентифицировать, то можно измерить и его «радиус». Это свойство полезно лишь в том случае, если рассматриваемые кластеры являются гиперсферами (т.е. имеют круглую форму) в многомерном пространстве, описываемом признаками.



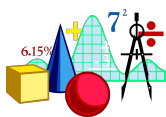
Свойства кластеров

Форма – это расположение точек в пространстве. Если кластеры имеют удлиненную форму, то вместо размера можно вычислить его «связность» - относительную меру расстояния между точками.



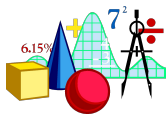
Свойства кластеров

Отделимость характеризует степень перекрытия кластеров и насколько далеко друг от друга они расположены в пространстве.

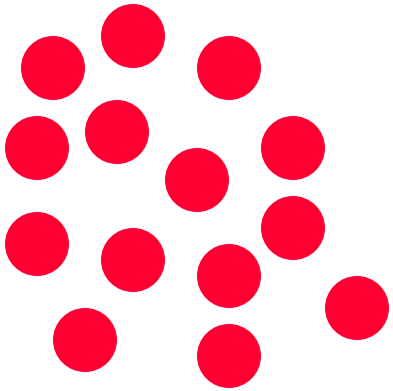


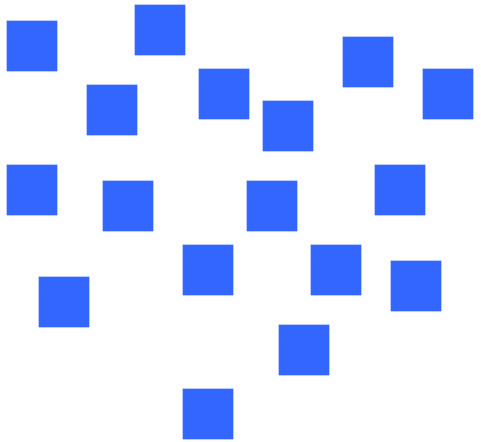
Кластерный анализ

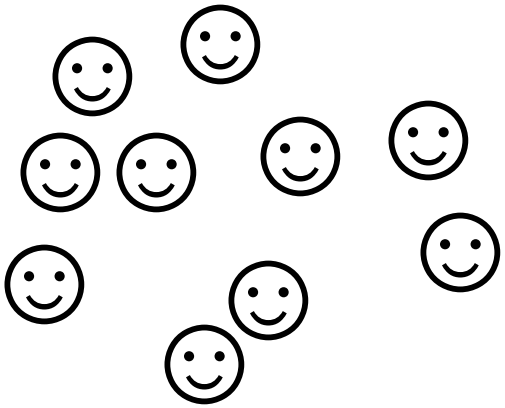
Таким образом, кластеры – это непрерывные области некоторого пространства с относительно высокой плотностью точек, отделенные от других таких же областей областями с относительно низкой плотностью точек.



Кластерный анализ



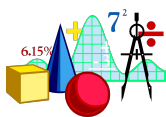




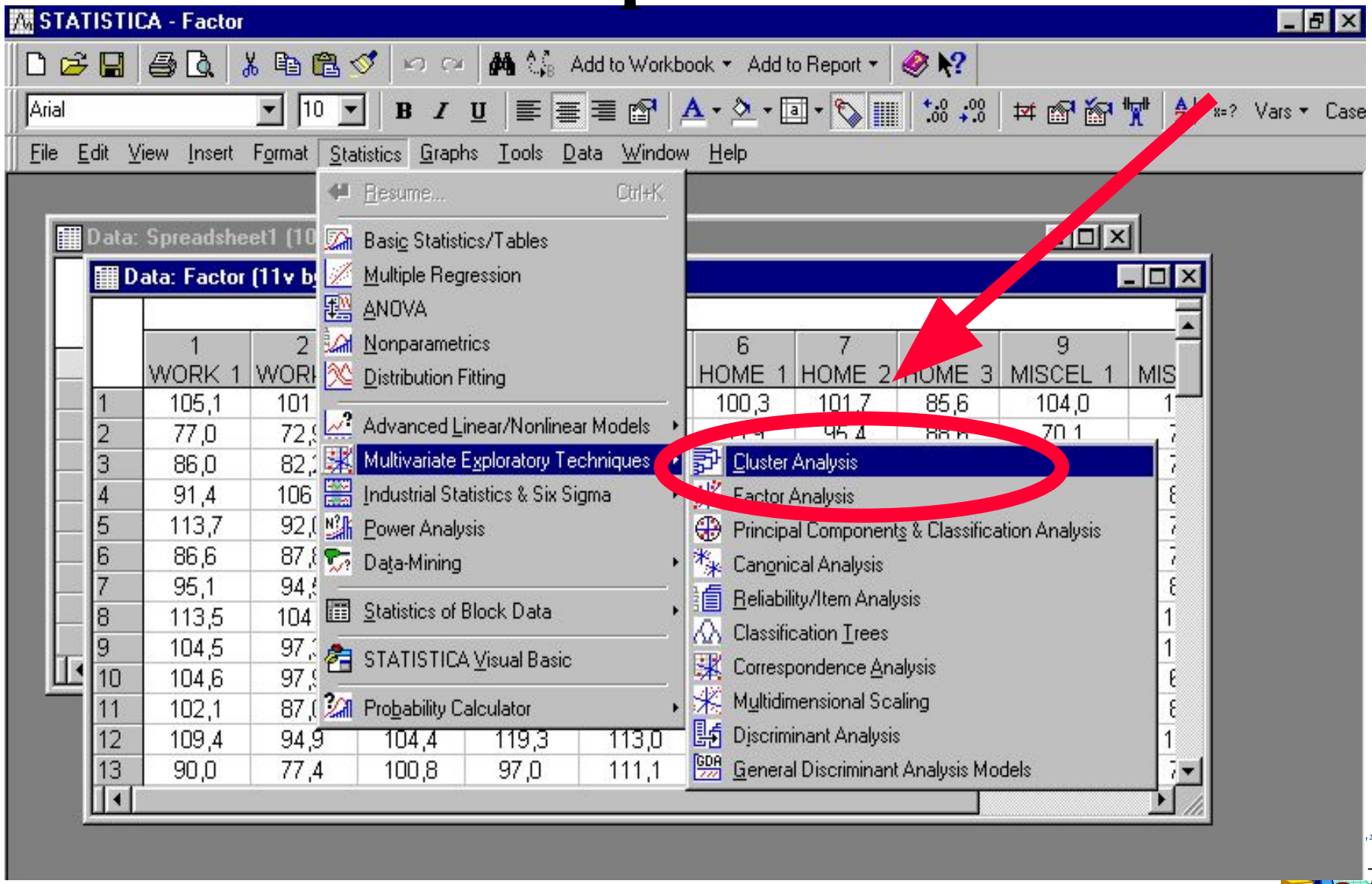
Кластерный анализ

можно сделать в программе STATISTICA,
в специальном модуле
Cluster Analysis

Statistics ⇒ Multivariate Exploratory Techniques
⇒ Cluster Analysis



Кластерный анализ



ПРЕДОСТЕРЕЖЕНИЯ!

1) Многие методы кластерного анализа – довольно простые процедуры, которые, как правило, не имеют достаточного статистического обоснования (то есть большинство методов являются эвристическими).

ПРЕДОСТЕРЕЖЕНИЯ!

2) Методы кластерного анализа разрабатывались для многих дисциплин, а потому несут на себе отпечатки специфики этих дисциплин.

ПРЕДОСТЕРЕЖЕНИЯ!

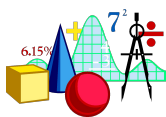
3) Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных.

ПРЕДОСТЕРЕЖЕНИЯ!

4) Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, и эта структура может не совпадать с искомой «реальной».

Выбор переменных

Основная проблема состоит в том, чтобы найти ту совокупность переменных, которая наилучшим образом отражает понятие сходства. В идеале переменные должны выбираться в соответствии с ясно сформулированной теорией, которая лежит в основе классификации.

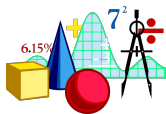


Выбор переменных - нормировка

Обычно при выполнении кластерного анализа данные подвергаются нормировке таким образом, чтобы среднее у всех переменных равнялось нулю, а дисперсия – единице.

Зачем?

Чтобы можно было сравнить все переменные между собой!



Выбор переменных - нормировка

$$Z = \frac{x_i - \bar{x}}{S}$$

где $\bar{x}$ – среднее значение показателя в группе;

x_i – значение показателя конкретного обследуемого;

S – стандартное отклонение;

Z – оценка индивидуального показателя.

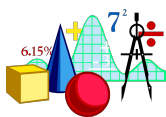
Выбор переменных - нормировка

В программе Statistica

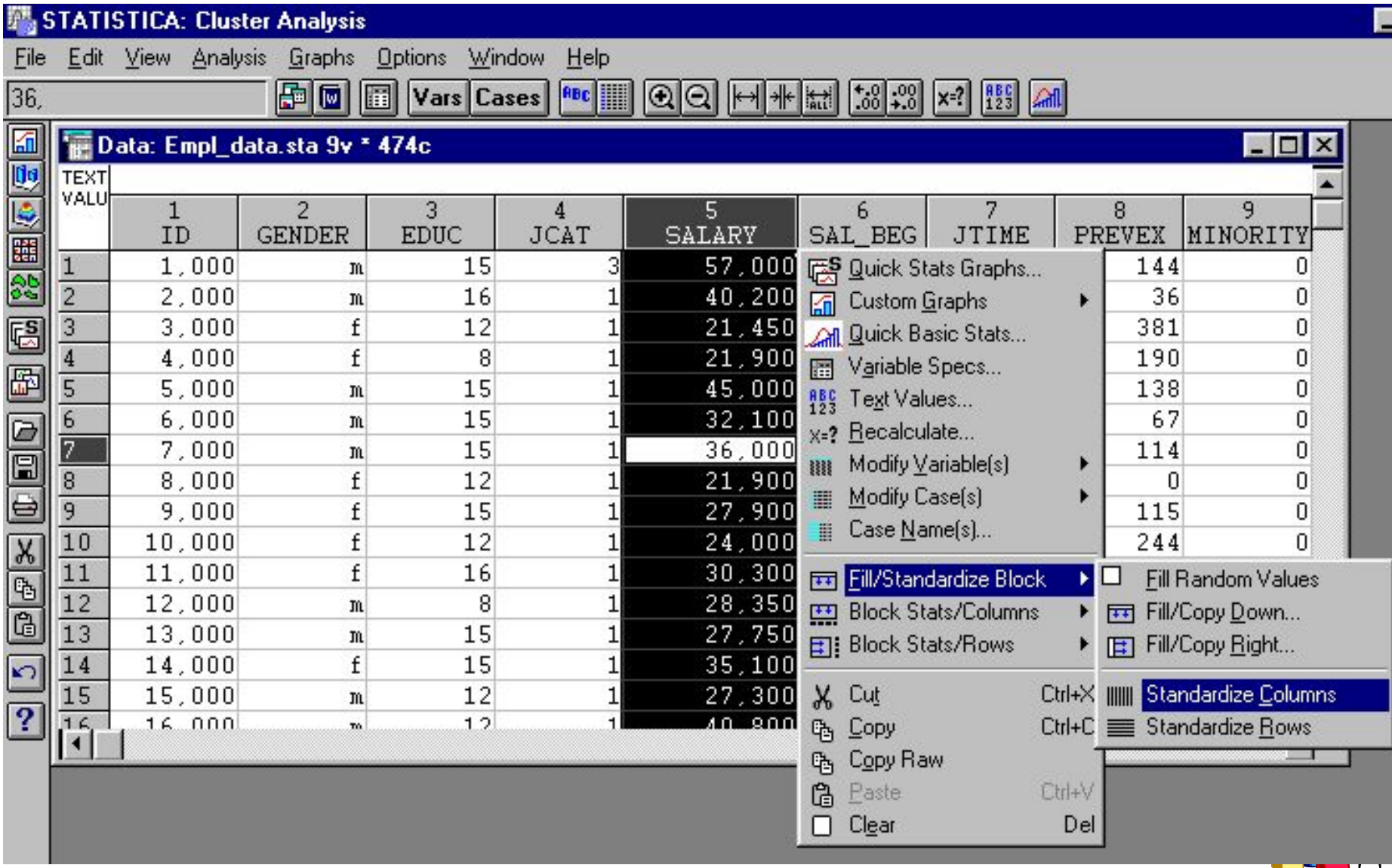
выбираем (выделяем) переменные, которые хотим нормировать,

затем нажимаем ПРАВУЮ кнопку мыши, и

Fill/Standardize Block → Standardize Columns…

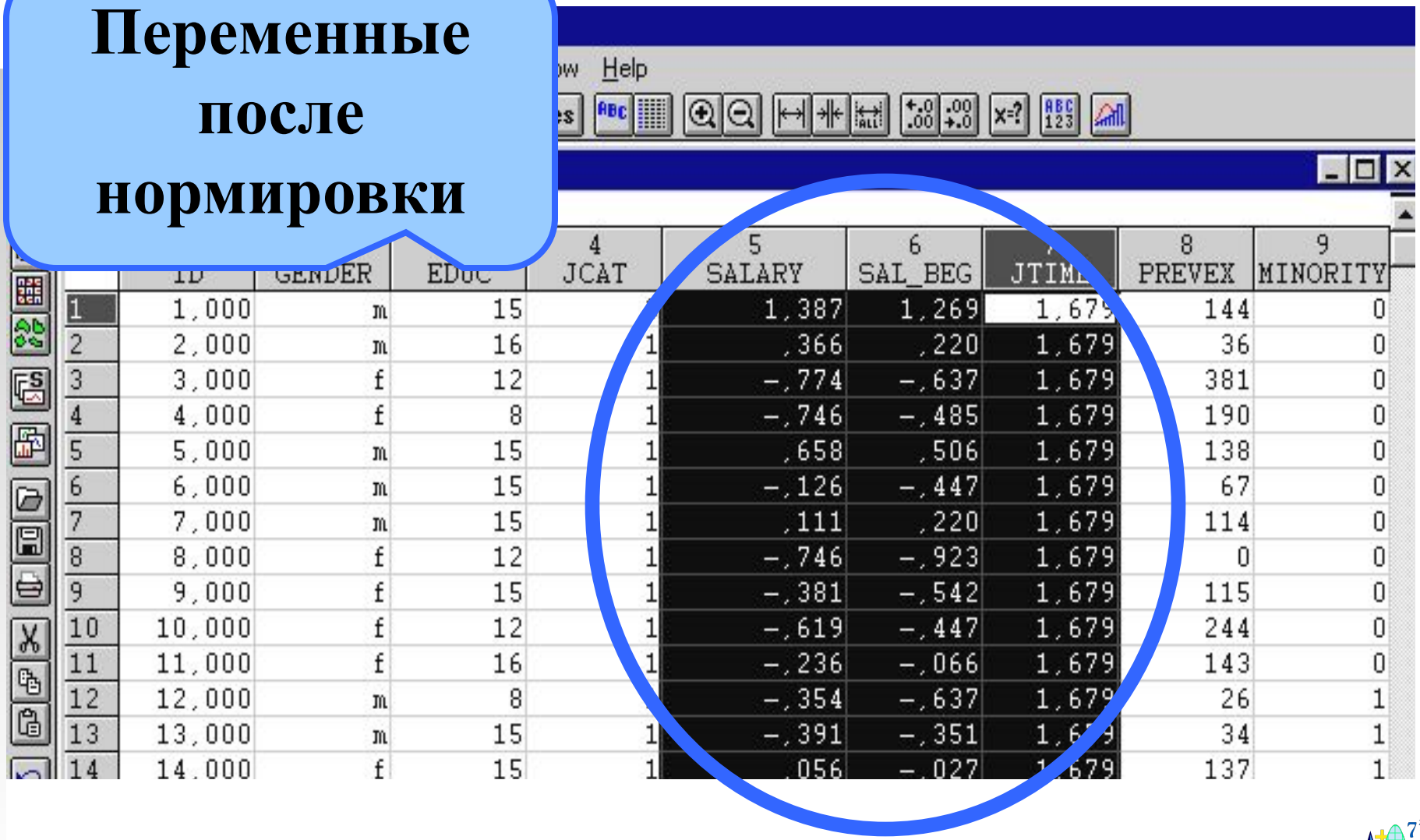


Выбор переменных - нормировка



Выбор переменных - нормировка

Переменные после нормировки

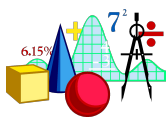


	ID	GENDER	EDUC	4 JCAT	5 SALARY	6 SAL_BEG	7 JTIME	8 PREVEX	9 MINORITY
1	1,000	m	15		1,387	1,269	1,679	144	0
2	2,000	m	16	1	,366	,220	1,679	36	0
3	3,000	f	12	1	-,774	-,637	1,679	381	0
4	4,000	f	8	1	-,746	-,485	1,679	190	0
5	5,000	m	15	1	,658	,506	1,679	138	0
6	6,000	m	15	1	-,126	-,447	1,679	67	0
7	7,000	m	15	1	,111	,220	1,679	114	0
8	8,000	f	12	1	-,746	-,923	1,679	0	0
9	9,000	f	15	1	-,381	-,542	1,679	115	0
10	10,000	f	12	1	-,619	-,447	1,679	244	0
11	11,000	f	16	1	-,236	-,066	1,679	143	0
12	12,000	m	8		-,354	-,637	1,679	26	1
13	13,000	m	15	1	-,391	-,351	1,679	34	1
14	14,000	f	15	1	,056	-,027	1,679	137	1

Выбор переменных - нормировка

Имеются, однако, некоторые разногласия относительно того, должна ли нормировка быть стандартной процедурой в кластерном анализе.

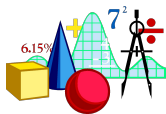
Нормировка к единичной дисперсии и нулевому среднему уменьшает различия между группами по тем переменным, по которым наилучшим образом обнаруживались групповые различия.



Выбор переменных - нормировка

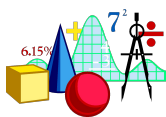
Более целесообразно проводить нормировку внутри групп (т.е. внутри кластеров), но, очевидно, этого нельзя сделать, пока объекты не разнесены по группам.

Гм ….



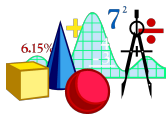
Выбор переменных - нормировка

Решение о проведении нормировки должно приниматься с учетом специфики решаемой задачи, при этом пользователь должен понимать, что результаты могут различаться в зависимости от принятого решения, хотя величина воздействия будет меняться от одного множества данных к другому.



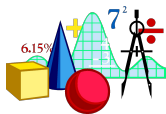
Выбор переменных - взвешивание

Взвешивание – это манипулирование значением переменной, позволяющее ей играть большую или меньшую роль в измерении сходства между объектами. Хотя эта идея и проста, ее практическое применение затруднительно. Видимо, имеет смысл взвешивать некоторые переменные априори, если для этого есть хорошее теоретическое обоснование.



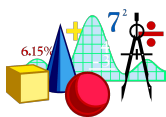
Методы кластерного анализа

Разные методы кластерного анализа соответствуют различным подходам к созданию групп, и применение различных методов к одним и тем же данным может привести к сильно различающимся результатам.



Методы кластерного анализа

Важно помнить, что выбранный метод должен находиться в согласии с ожидаемым характером классификации, применяемыми признаками и мерой сходства.

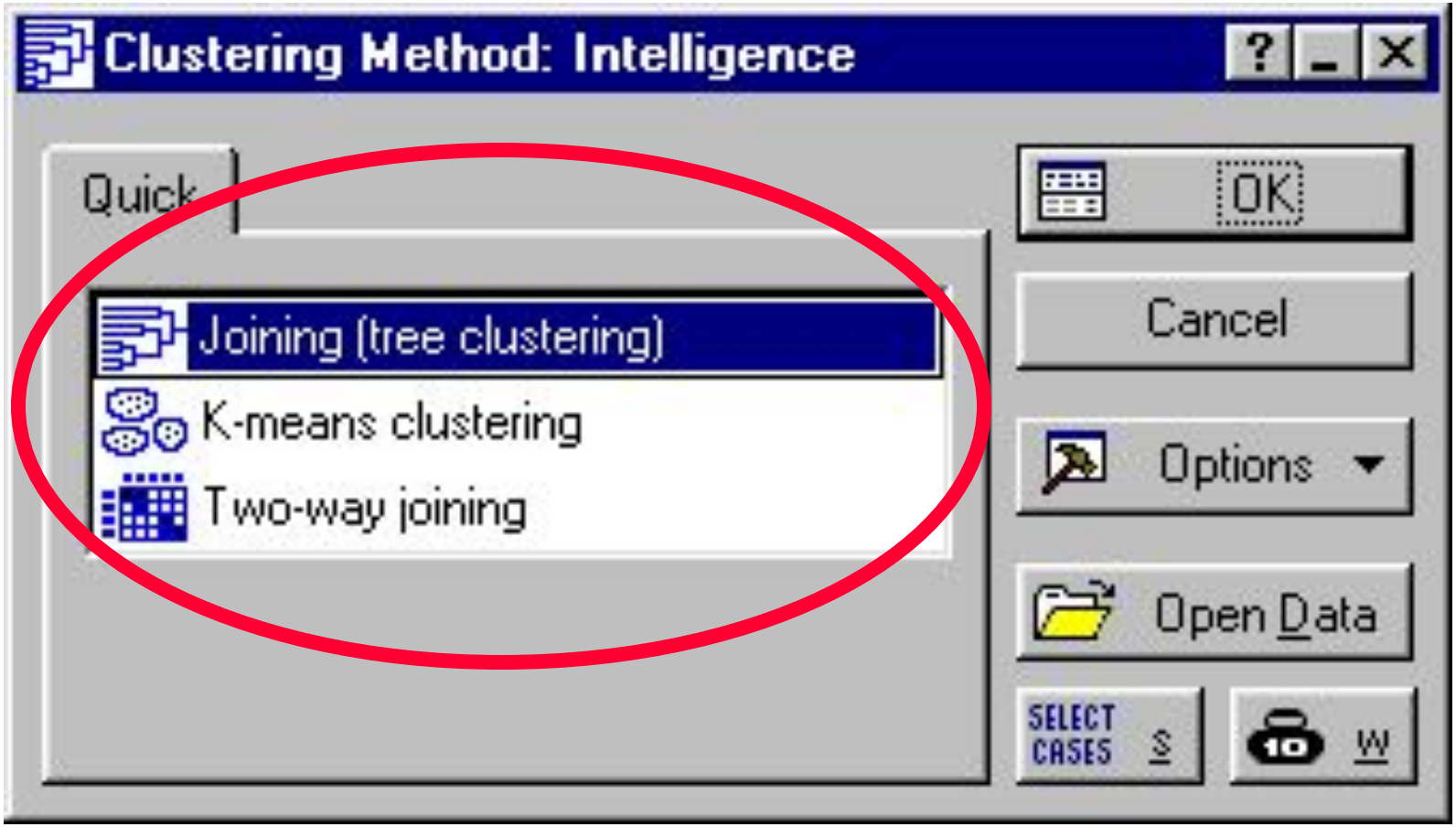


Методы кластерного анализа

В программе STATISTICA реализованы следующие методы кластеризации:

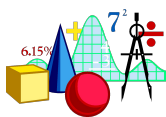
- **иерархический агломеративный (объединительный) метод – joining (tree clustering),**
- **итеративный метод k-средних (k-means clustering)**
- **двухвходовое объединение (two-way joining).**

Методы кластерного анализа

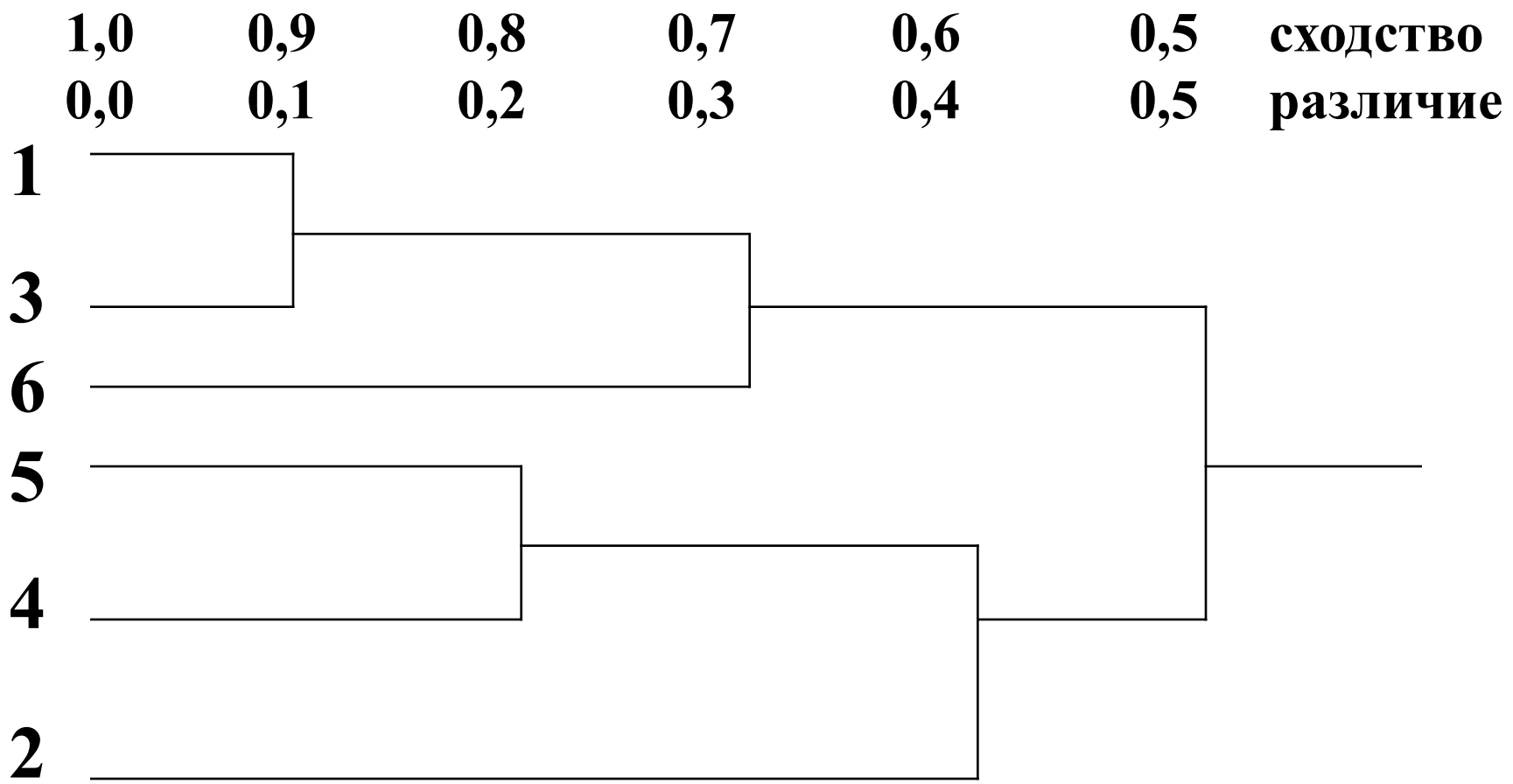


Агломеративный метод

В агломеративных методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде дендрограммы, или дерева объединения.



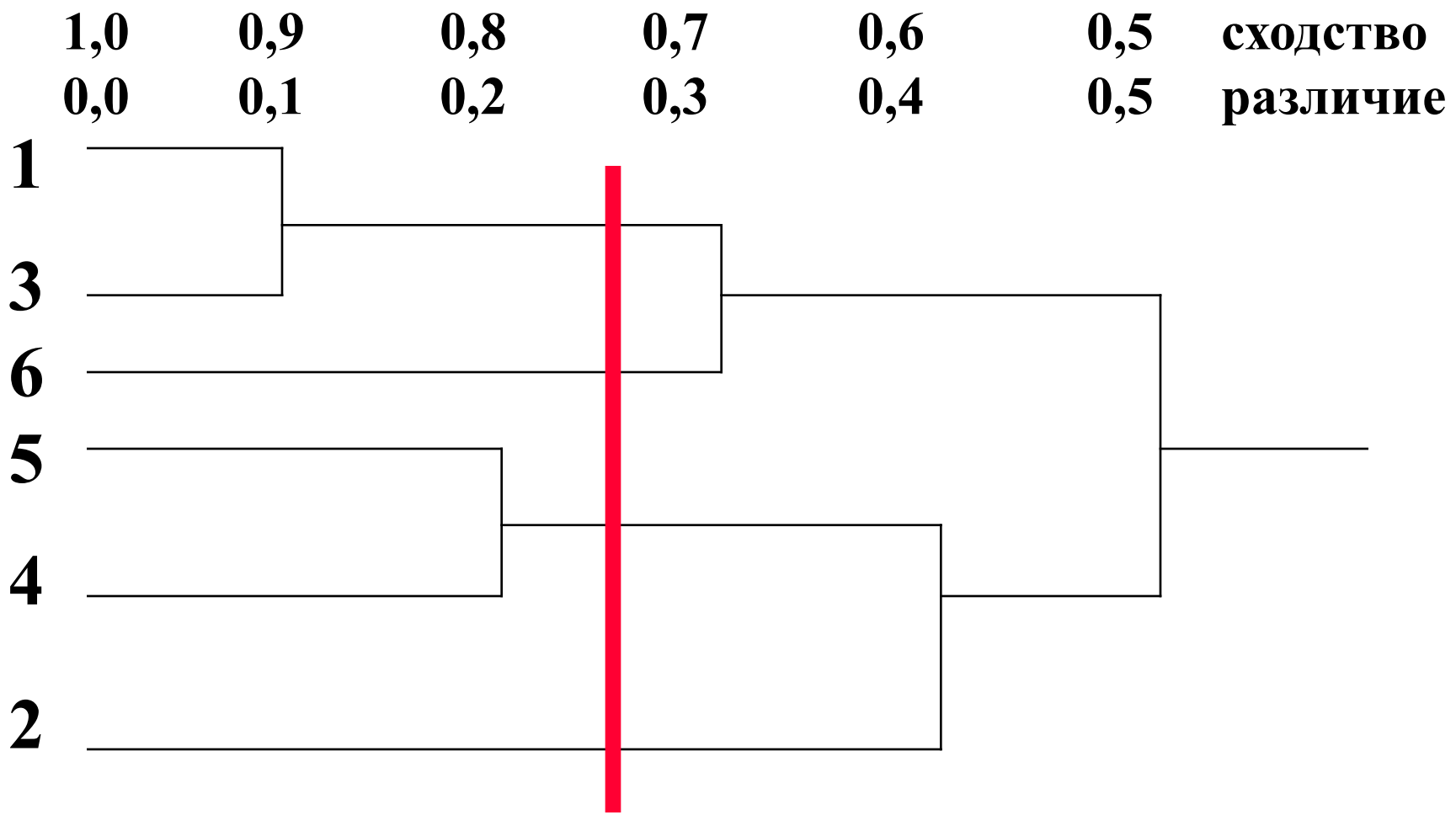
Агломеративный метод



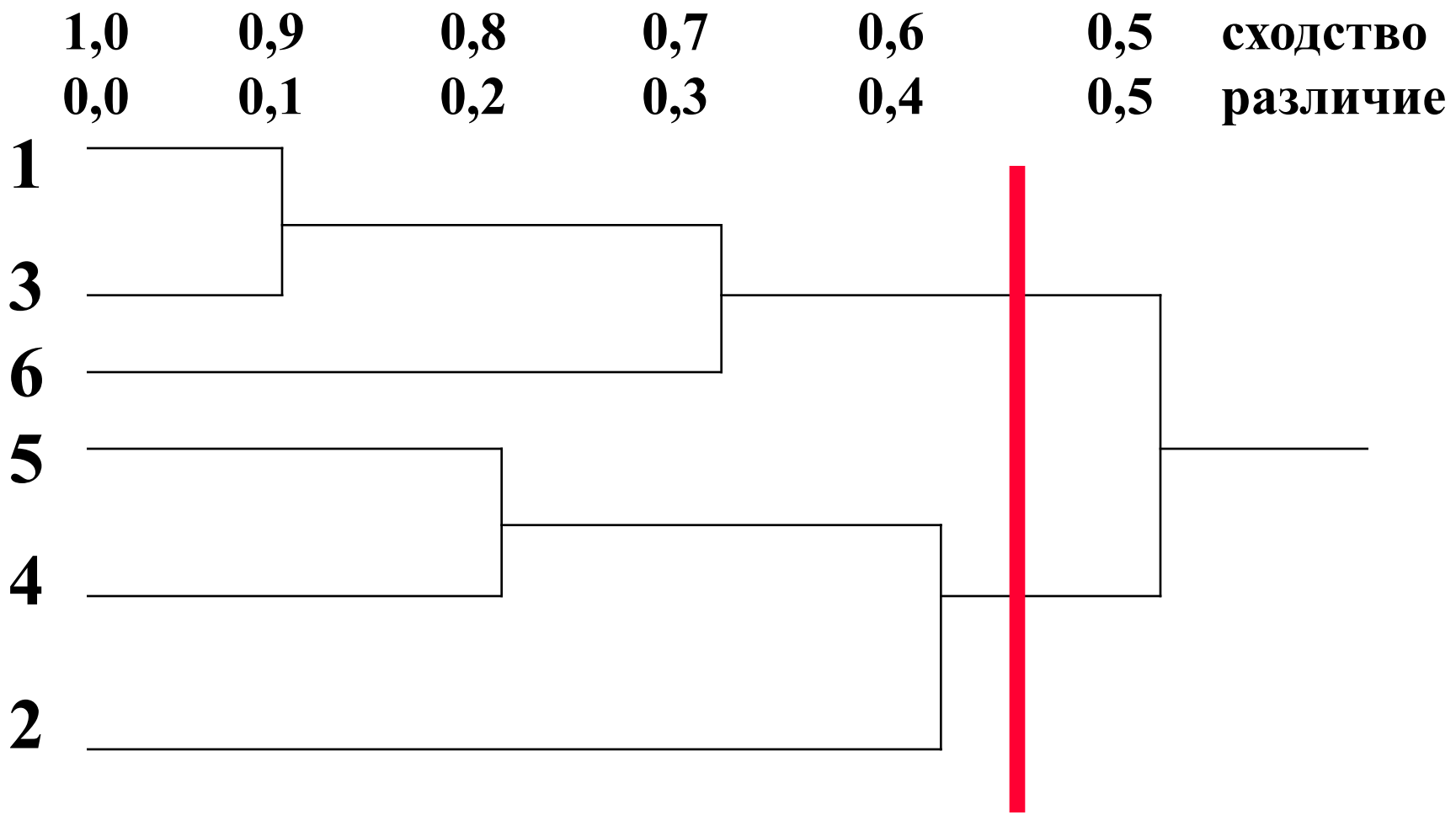
Агломеративный метод

Рубить дерево можно в любом месте!

Агломеративный метод

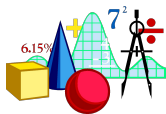


Агломеративный метод



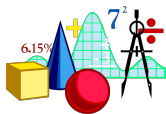
Меры сходства

Количественное оценивание сходства отталкивается от понятия метрики или расстояния (distance) между объектами. Интуитивно понятно, что чем меньше расстояние между объектами, тем больше сходство между ними.



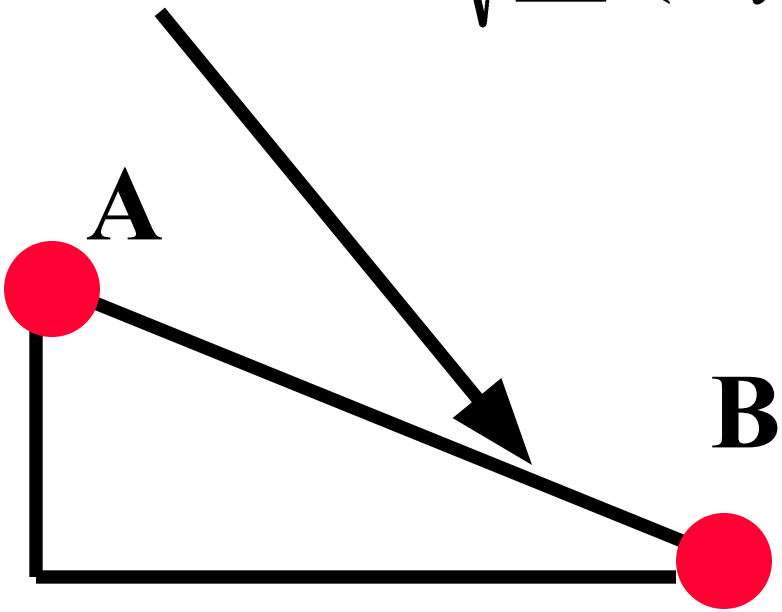
Меры сходства

☺ **Евклидова метрика – наиболее часто используемая мера сходства. Вы просто возводите в квадрат расстояния по каждой координате, суммируете их и из полученной суммы извлекаете квадратный корень.**



Меры сходства

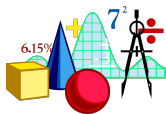
Расстояние (x,y)= $\sqrt{\sum (x_i - y_i)^2}$



Меры сходства

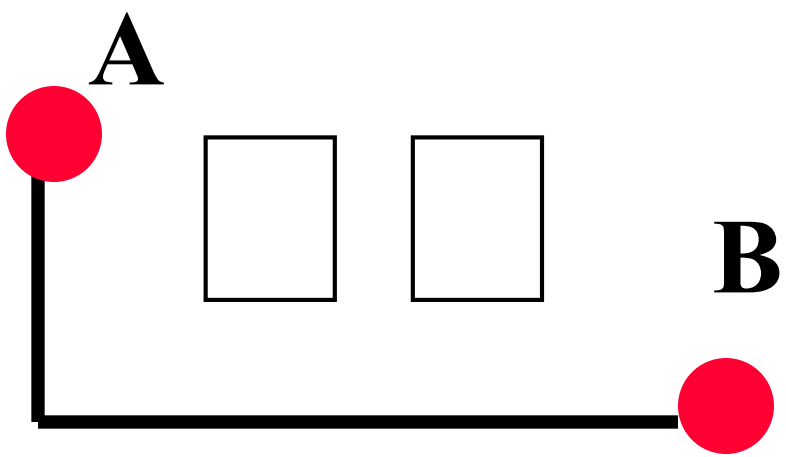
☺ **Квадрат евклидовой метрики.**

Расстояние (x,y)= $\sum (x_i - y_i)^2$



Меры сходства

☺ **Манхэттенское расстояние, или «расстояние городских кварталов». В этом случае просто берутся абсолютные значения покоординатных расстояний и суммируются.**



Меры сходства

Аналогия в декартовой плоскости приводит к перемещениям только по линиям, параллельным осям координат, и соответственно, к манхэттенскому расстоянию.

Расстояние (x,y)= $\sum \left| x_i - y_i \right|$

Меры сходства

- **Метрика Чебышева**

Расстояние (x,y)= $max\left|x_i - y_i\right|$

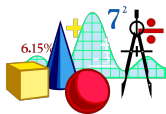
Меры сходства

☺ **Метрика Минковского.**

Расстояние (x,y)= $\sqrt[r]{\sum (x_i - y_i)^p}$

Меры сходства

☺ **Коэффициент корреляции Пирсона (точнее, 1 - коэффициент корреляции Пирсона)**



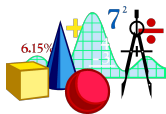
Меры сходства

☺ **Коэффициент совстречаемости – метрика, наиболее пригодная для данных, представленных в шкалах наименований. Вычисляется как**

Расстояние (x,y)= $\left(число\ x_i \neq y_i\right)/i$

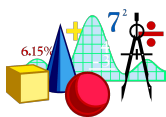
Меры сходства

Однозначного ответа на вопрос, какую из мер сходства выбрать, не существует. Ответ зависит от типа данных и природы решаемой задачи.



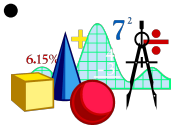
Правила объединения

Кроме выбора меры сходства, исследователю предстоит задача выбора правила иерархического объединения кластеров. В программе реализованы следующие методы:



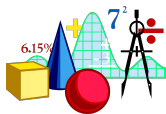
Правила объединения

Single linkage **– метод одиночной связи. На первом шаге объединяются два объекта, имеющие между собой максимальную меру сходства. На следующем шаге к ним присоединяется объект с максимальной мерой сходства с *одним* из объектов кластера. Таким образом процесс продолжается дальше. Для включения объекта в кластер требуется максимальное сходство лишь с одним членом кластера.**



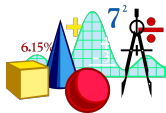
Правила объединения

Complete linkage **– метод полной связи. Этот метод позволяет устранить указанный недостаток. Здесь мера сходства между объектом – кандидатом на включение в кластер и всеми членами кластера не может быть меньше некоторого порогового значения.**



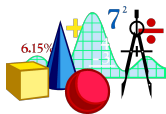
Правила объединения

Unweighted pair group average **–метод «средней связи». В этом методе вычисляется среднее сходство рассматриваемого объекта со всеми объектами в уже существующем кластере, а затем, если найденное среднее значение сходства достигает или превосходит некоторый заданный пороговый уровень сходства, объект присоединяется к этому кластеру. Чаще всего берется просто среднее арифметическое мер сходства между объектами кластера и кандидатом на включение.**



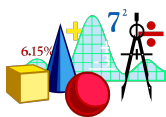
Правила объединения

Weighted pair group average – **взвешенный метод «средней связи». Аналогичен предыдущему, за исключением того, что в данном случае в качестве весов берутся размеры соответствующих кластеров (т.е., число объектов в кластере). Этот метод лучше использовать, если есть подозрения, что кластеры будут иметь размеры, сильно различающиеся между собой.**



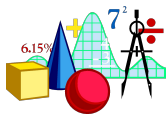
Правила объединения

Unweighted pair group centroid – **центроидный метод. Расстояние между двумя кластерами определяется как евклидово расстояние между центрами (средними) этих кластеров. Кластеризация осуществляется поэтапно: на каждом шаге объединяют два кластера, расстояние между которыми минимально.**



Правила объединения

Weighted pair group centroid –
взвешенный центроидный метод. Аналогичен предыдущему, за исключением того, что в данном случае в качестве весов берутся размеры соответствующих кластеров (т.е., число объектов в кластере).



Правила объединения

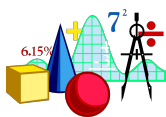
Ward method – метод Уорда. Идея этого метода состоит в том, чтобы проводить объ… …льное при… …суммы квад… …ений, т… оптимизировать минима… дисперсию внутри клас…

Это хороший метод!

Метод k-средних

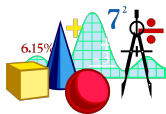
Это итеративный метод, который работает непосредственно с объектами, а не с матрицей сходства.

Он отличается тем, что позволяет *заранее* задать число кластеров. Это число определяет сам пользователь, исходя из имеющейся задачи и предсказаний теории.



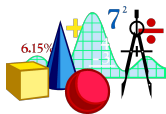
Метод k-средних

Метод k-средних разобьет все объекты на заданное количество кластеров, которые будут максимально различаться между собой.



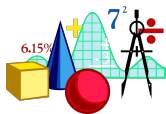
Метод k-средних

В этом методе объект относится к тому классу, расстояние до которого минимально. Расстояние понимается как евклидово расстояние, то есть объекты рассматриваются как точки евклидова пространства.



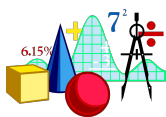
Метод k-средних

Вначале задается некоторое разбиение данных на кластеры (число кластеров определяется пользователем) и вычисляются центры тяжести кластеров. Затем происходит перемещение каждой точки в ближайшей к ней кластер.



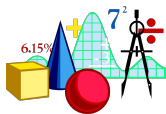
Метод k-средних

Затем снова вычисляются центры тяжести новых кластеров и процесс повторяется, пока не будет найдена стабильная конфигурация (то есть кластеры перестанут изменяться) или число итераций не превысит заданное пользователем.



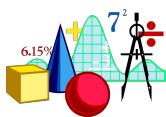
Метод k-средних

Можно сказать, что вычислительная процедура данного метода представляет собой дисперсионный анализ «наоборот». Программа начинает работу с k случайных кластеров, а затем перемещает объекты из одного кластера в другой с целью (1) минимизировать вариативность (дисперсию) внутри кластера и (2) максимизировать вариативность между кластерами.



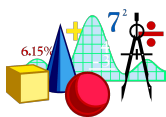
Метод k-средних

Это аналогично дисперсионному анализу «наоборот» в том смысле, что в дисперсионном анализе при определении значимости различий в средних значениях групп оценивается межгрупповая дисперсия в сравнении с внутригрупповой дисперсией.



Метод k-средних

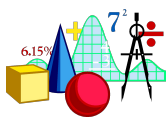
В методе k-средних программа пытается перемещать объекты между группами (кластерами) таким образом, чтобы получить наиболее значимые результаты дисперсионного анализа. Поэтому и результаты этого самого дисперсионного анализа приводятся в разделе результатов применения данного метода.



Метод k-средних

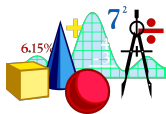
Кроме числа кластеров, пользователю также необходимо выбрать условие, которое задает начальные центры кластеров. Существует три возможности:

- ***Maximize between-cluster distances.***
- ***Sort distances and take observations at constant intervals.***
- ***Choose the first N (number of clusters) clusters observations.***



Maximize between-cluster distances

Если выбрано это условие, то за центр кластера принимается наблюдение или объект, а выбор объектов следует правилу максимизации начальных расстояний между кластерами.

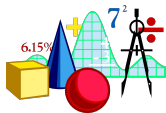


Maximize between-cluster distances

В этом случае программа

(1) выберет сначала первые N (число кластеров, заданное вами) наблюдений в качестве центров кластеров;

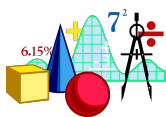
(2) последующие наблюдения заменят выбранные центры кластеров, если наименьшее расстояние от них до любого другого центра кластера больше, чем наименьшее расстояние между кластерами.



Sort distances and take observations at constant intervals

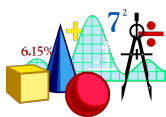
Если выбрано это условие, расстояния между объектами сначала будут упорядочены, а затем объекты с одинаковыми расстояниями будут выбраны в качестве центров кластеров.

(Выбирается по умолчанию)



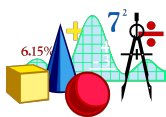
Choose the first N (number of clusters) clusters observations

При выборе этого условия первые N (количество кластеров) наблюдений будут выбраны в качестве начальных центров кластеров. Таким образом, это условие дает пользователю возможность контроля выбора начальной конфигурации. Это бывает полезно, если исследователь хочет проверить какие-то начальные предположения о составе кластеров. В этом случае передвиньте те наблюдения, вокруг которых вы хотите сгруппировать все остальные, в начало файла.



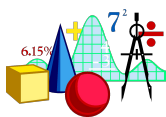
Two-way joining

применяется в тех (сравнительно редких) случаях, когда исследователь полагает, что и переменные, и наблюдения одновременно вносят вклад в определение «реальной» структуры. Результаты этого метода достаточно сложно интерпретировать, так как сходство между различными кластерами может объясняться различными подмножествами переменных, что приводит к неоднородности результирующей структуры.



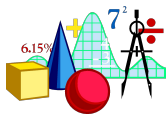
Алгоритм кластерного анализа

1. **Заносим данные в программу. Возможно проводим процедуру нормировки.**
2. **Выбираем метод - агломеративный (объединительный) метод (joining (tree clustering)), итеративный метод k-средних (k-means clustering) или двухвходовое объединение (two-way joining).**



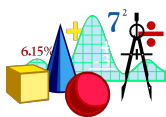
Алгоритм кластерного анализа

3. **Если выбран метод tree clustering, то выбираем метод объединения объектов в кластеры.**
4. **Затем выбираем правило определения сходства.**
5. **Если полученная структура не устраивает исследователя по каким-то параметрам или не поддается осмысленной интерпретации, то пробуем другие правила определения сходства (возвращаемся на п. 4).**



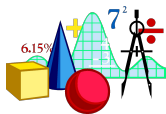
Алгоритм кластерного анализа

6. **Если ничего не получается, то можно попробовать разные методы объединения объектов в кластеры (возвращаемся на п.3).**
7. **Если это ничего не дает, то можно попробовать другой метод кластеризации (возвращаемся на п. 2)**



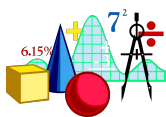
Алгоритм кластерного анализа

8. **Если выбран метод k-средних (k-means clustering), то выбираем число кластеров.**
9. **Затем выбираем условие, которое задает начальные центры кластеров.**
10. **Задаем минимальное число итераций побольше.**
11. **Если результаты не нравятся, можно попробовать другое условие для вычисления начальных центров (возвращаемся на п. 9).**



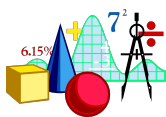
Алгоритм кластерного анализа

12. **Если и это ничего не дает, то можно попробовать взять другое количество кластеров (возвращаемся на п. 8).**
13. **Если это ничего не дает, то можно попробовать другой метод кластеризации (возвращаемся на п. 2)**



Алгоритм кластерного анализа

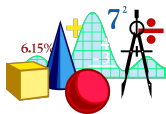
14. **Если выбран метод two-way joining, то возможности изменить что-либо, кроме переменных, участвующих в анализе, у пользователя нет. Поэтому следует просто попытаться интерпретировать результаты. Если это не получается, то, видимо, вы выбрали неудачный метод, и следует вернуться на п. 2.**



Полезная литература

Просто и доходчиво кластерный анализ изложен в

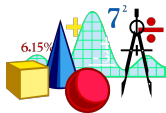
☺ **Боровиков В. Программа *STATISTICA для студентов и инженеров*. – Компьютер Пресс: Москва – 2001. – 301 с.**



Полезная литература

Более подробное описание можно найти в книге:

☺ **Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика**

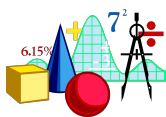


Пример

Цели дипломной работы:

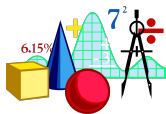
1) выделить группы подростков, характеризующиеся различными предпочтениями жанров киноискусства и телепередач

2) изучить взаимосвязь агрессивности подростков с передачами и фильмами, которые они любят и смотрят регулярно



Пример

Попытаемся разделить учащихся на основании сразу нескольких критериев, т.е. всех перечисленных жанров киноискусства и телепередач, а для решения этой задачи используем кластерный анализ (метод k-средних).



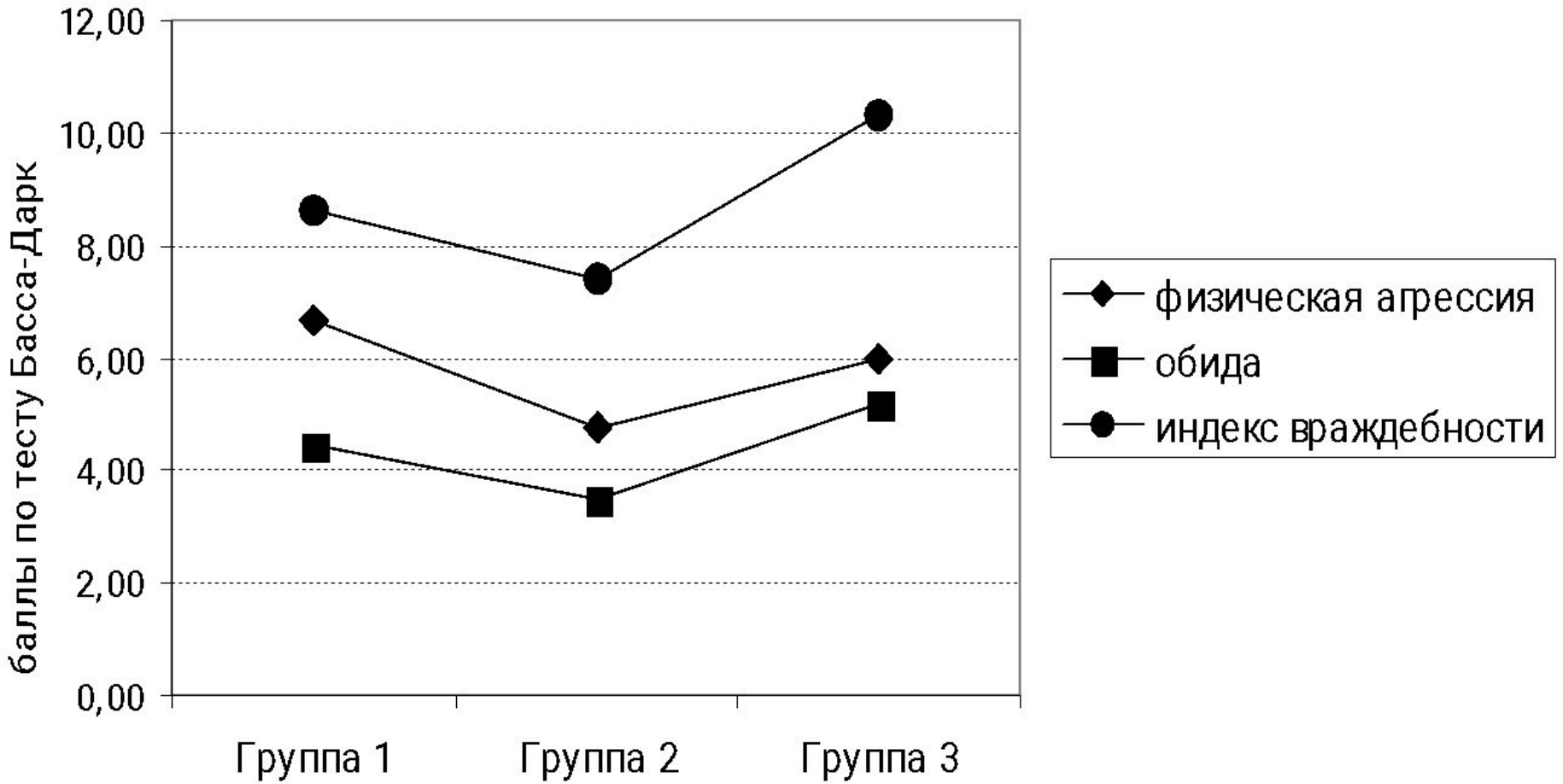
Пример

	Кластер 1	Кластер 2	Кластер 3
Комедии	0,89	0,89	0,83
Боевики	0,42	0,00	0,07
Мелодрамы	0,00	1,00	0,67
Фильмы ужасов	0,11	0,33	0,50
…			
Телепередачи			
Спортивные	0,15	0,22	0,67
Музыкальные	0,56	0,78	0,08
Развлекательные	1,00	0,83	0,00
…			

Пример

	df_1	df_1	F	p
Физическая агрессия	**2**	**47**	**6,68**	**0,00**
Косвенная агрессия	**2**	**47**	**0,61**	**0,55**
Раздражение	**2**	**47**	**0,06**	**0,94**
Негативизм	**2**	**47**	**0,02**	**0,98**
Обида	**2**	**47**	**2,90**	**0,04**
…				
Индекс агрессивности	**2**	**47**	**0,70**	**0,50**
Индекс враждебности	**2**	**47**	**4,85**	**0,02**

Пример


12,00
10,00
8,00
6,00
4,00
2,00
0,00
баллы по тесту Басса-Дарк
Группа 1
Группа 2
Группа 3
физическая агрессия
обида
индекс враждебности

Пример

Таблица X

Уровни статистической значимости апостериорного критерия Дункана для сравнения выраженности физической агрессивности у трех групп испытуемых

	Группа 1	Группа 2
Группа 1	…	…
Группа 2	**0,0035**	…
Группа 3	0,2723	**0,0427**

Неплохо и перекусить!