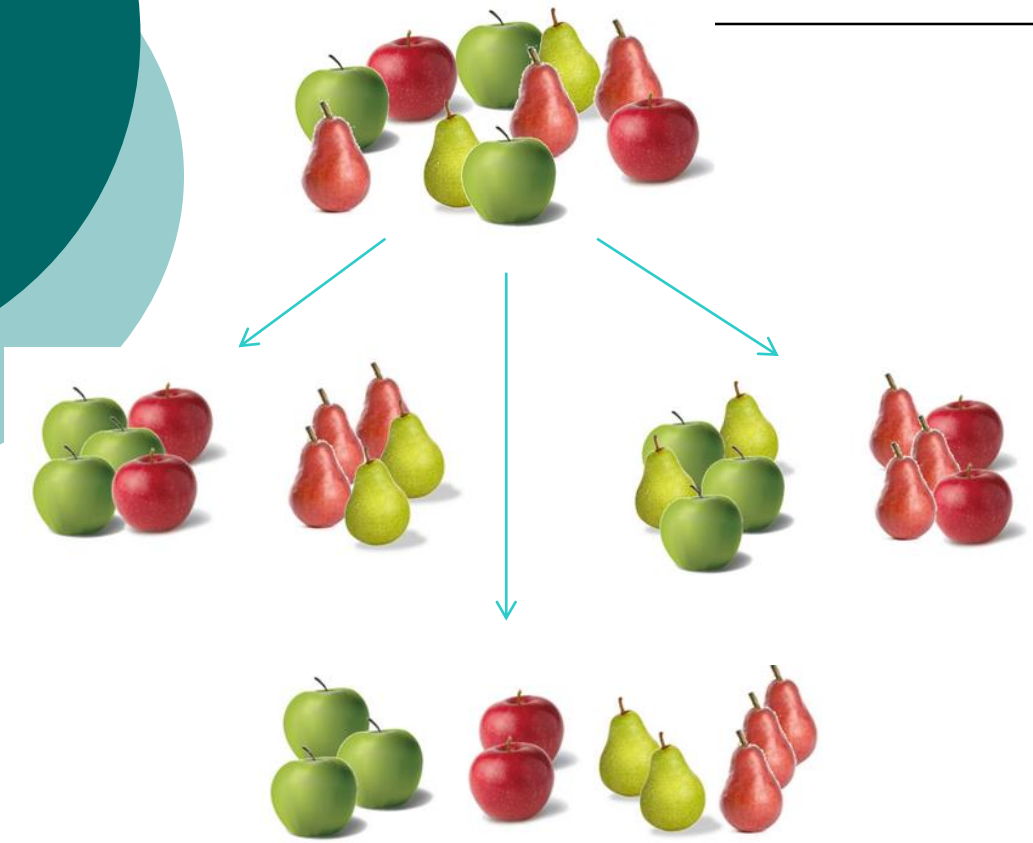
Кластерный анализ данных

Лекция 4.

Основные определения и понятия

- **Кластерный анализ (КА) – общее название множества вычислительных процедур, используемых при создании классификации.**
- **Многомерная статистическая процедура, выполняющая обработку данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы, называемые кластерами.**

Основные определения и понятия



- Кластеризация (таксономия) – процесс выделения групп схожих объектов их исходного множества.
- Применяем различные методы - получаем разбиения на кластеры

Постановка задачи КА

***Дано*:**

Множество объектов данных X=(X1, X2,…,Xn)

Для каждого объекта измерено p признаков:

$$X_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$$

Исходные данные

номер	возраст	стаж	кол-во часов	доход	профессия	пол
X1	32	1,6	13,9	9,5	1	2
X2	40	0,4	6	8,7	0	1
X3	30	2,5	24,6	8,1	1	2
X4	36	0,4	6,5	7,8	0	1
X5	41	0,3	7,1	10,5	0	2
X6	39	0,7	6,7	10,2	0	1
X7	24	5	76,2	13,3	2	2
X8	30	5	24,1	10,2	1	2

Постановка задачи КА

Необходимо: **Построить множество кластеров**

$$C=\{c1, c2,\ldots, cm,\ldots, cq\}$$

и отображение F множества X на множество C, т.е.

$$F: X->C.$$

Отображение F задает модель данных, являющуюся решением задачи.

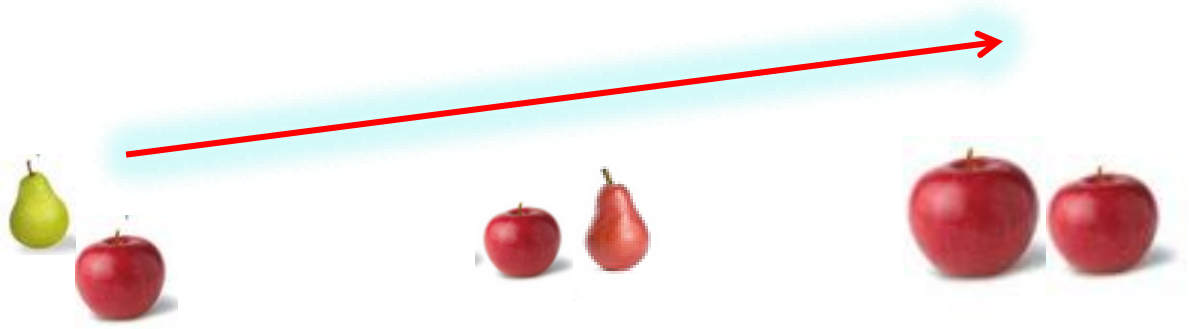
$$c_m = \{X_i, X_k \mid X_i \in X, X_k \in X, d(X_i, X_k) \le \varepsilon\}$$

ε **- величина, определяющая меру близости для включения объектов в один кластер**

$d(X_i, X_k)$ **- мера близости, расстояние**

Сходство объектов. Мера сходства

- Мера сходства (мера близости) — величина, имеющая предел и возрастающая с увеличением близости объектов.



Постановка задачи КА

При этом обучающая выборка отсутствует;

априорная информация о характере распределения

измерений внутри групп отсутствует. Качество решения задачи определяется количеством верно классифицированных

объектов.

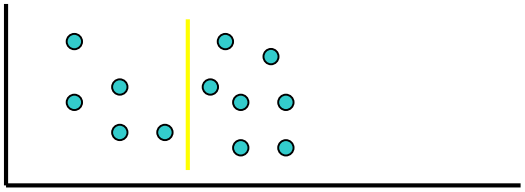
Результат: **группы (кластеры, таксоны).**

Задачи кластеризации

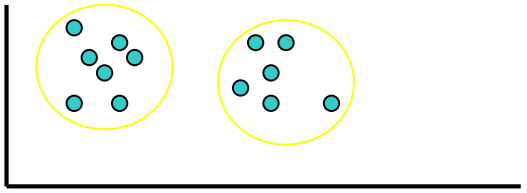
- **Кластер-анализ еще называют численной таксономией; распознаванием образов с самообучением; классификацией без обучения.**

Принципиально решается две разные задачи классификации в ходе КА:

1. **ТИПИЗАЦИЯ** (исследуемую совокупность следует разбить на сравнительно небольшое число групп – аналог интервалов группирования при обработке одномерных наблюдений).



2. **КЛАССИФИКАЦИЯ** (естественное расслоение на четко выраженные кластеры). Во второй постановке задача не всегда имеет решение.



Методологические этапы проведения кластерного анализа:

1. **Выбор переменных (признаков) на основе которых исходная совокупность объектов разбивается на кластеры;**
2. **Выбор меры сходства (меры расстояний) между объектами;**
3. **Выбор метода (процедуры) КА;**
4. **Проведение кластеризации;**
5. **Оценка полученных результатов.**

Выбор переменных (признаков)

- ***нормировка переменных***
- ***преобразование данных***
- ***взвешивание переменных***

Выбор меры сходства.

Кластеризация объектов проводится на основе вычисления меры сходства между объектами.

Меры сходства подразделяются на четыре группы:

- **коэффициенты корреляции;**
- **меры расстояния;**
- **коэффициенты ассоциативности;**
- **вероятностные коэффициенты сходства.**

Выбор меры сходства.

Меры сходства должны удовлетворять 4 критериям, чтобы быть метрикой

1. Симметрия. Для объектов X,Y расстояние между ними удовлетворяет условию:

$$d(X,Y)=d(Y,X)\geq 0$$

2. Неравенство треугольника: Для объектов X,Y,Z расстояние между ними удовлетворяет условию: длина любой стороны треугольника меньше или равна сумме двух других сторон.

$$d(X,Y)\leq d(X,Z)+d(Y,Z)$$

3. Различимость нетождественных объектов: Для объектов X,Y если $d(X,Y)\neq 0 \Rightarrow X\neq Y$
4. Неразличимость идентичных объектов. Для двух идентичных объектов X,X*:

$$d(X,X^*)=0$$

Коэффициент корреляции

$$r_{jk} = \frac{\sum_{i=1}^{p} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{p} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{p} (x_{ik} - \bar{x}_k)^2}}$$

x_{ij} - значение *i*-ой переменной (признака) для *j*-го объекта;

$\bar{x}_j$ - среднее всех значений переменных для *j*-го объекта;

p – число переменных.

Коэффициент корреляции

$$r_{jk} = \frac{\sum_{i=1}^{p} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{p} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{p} (x_{ik} - \bar{x}_k)^2}}$$

номер	возраст	стаж	кол-во часов	доход
X1	33	1	13	9
X2	40	4	6	10

Коэффициент корреляции

$$r_{jk} = \frac{\sum_{i=1}^{p} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{p} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{p} (x_{ik} - \bar{x}_k)^2}}$$

номер	возраст	стаж	кол-во часов	доход
X1	33	1	13	9
X2	40	4	6	10

$$r_{x1x2} = 0{,}95$$

Меры расстояния.

Два объекта идентичны, если описывающие их переменные принимают одинаковые значения, в этом случае расстояние между ними равно нулю.

Евклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

,где x_{ik} - значение к-ой переменной для j-го объекта.

Вычисляется в Statistica по не нормированным данным, поэтому в случае сильно различающихся шкал измерения переменных необходимо быть осторожнее.

- ***Квадрат евклидова расстояния***

Для того чтобы избежать квадратного корня: d_{ij}^2

Меры расстояния.

Евклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

, где x_{ik} - значение к-ой переменной для j-го объекта.

номер	возраст	стаж	кол-во часов	доход
X1	33	1	13	9
X2	40	4	6	10
Xmax	41	5	76	13

Меры расстояния.

Евклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

, где x_{ik} - значение к-ой переменной для j-го объекта.

номер	возраст нормир.	стаж нормир.	кол-во часов нормир.	доход нормир.
X1	0,8	0,2	0,17	0,69
X2	0,975	0,8	0,08	0,77
Xmax	41	5	76	13

Меры расстояния.

Евклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

,где x_{ik} - значение к-ой переменной для j-го объекта.

номер	возраст нормир.	стаж нормир.	кол-во часов нормир.	доход нормир.
X1	0,8	0,2	0,17	0,69
X2	0,975	0,8	0,08	0,77
Xmax	41	5	76	13

$$d_{x1x2} = 0,63$$

Меры расстояния.

Класс метрик Минковского

(в Statistica называется степенным расстоянием)

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{p} \right)^{1/r}$$

Если p=r=2, то евклидово расстояние, p и r задают пользователи. Параметр p взвешивает расстояние между двумя переменными, а r – расстояние между двумя объектами.

Меры расстояния.

- *Расстояние городских кварталов*

(манхэттенское расстояние)

Это расстояние является просто суммой разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

Манхэттенское расстояние вычисляется по формуле:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Меры расстояния.

- *Расстояние городских кварталов*

(манхэттенское расстояние)

Это расстояние является просто суммой разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

Манхэттенское расстояние вычисляется по формуле:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

$$d_{x1x2} = 0{,}945$$

Меры расстояния.

- ***Расстояние Чебышева.***

Это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате (каким-либо одним измерением).

Расстояние Чебышева вычисляется по формуле:

$$d_{ij} = \max|x_{ik} - x_{jk}|$$

Меры расстояния.

- ***Расстояние Чебышева.***

Это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате (каким-либо одним измерением).

Расстояние Чебышева вычисляется по формуле:

$$d_{ij} = \max|x_{ik} - x_{jk}|$$

$$d_{x1x2} = 0,6$$

Меры расстояния.

- ***Пиковое расстояние***

$$d_{ik} = 1/p \sum_{j=1}^{p} \frac{|x_{ij} - x_{kj}|}{x_{ij} + x_{kj}}$$

Выбор меры можно обосновать только в случае содержательного анализа изучаемых объектов.

- ***Процент несогласия.***

Эта мера используется в тех случаях, когда данные являются категориальными.

Это расстояние вычисляется по формуле:

$$d_{ij} = number\, x_{ik} \neq x_{jk}/p$$

Коэффициенты ассоциативности

Используются при бинарных переменных (предложено более 30 таких коэффициентов).

Таблица ~~ассоциативности~~

	1	0
1	a	b
0	c	d

Самые распространенные:

- **коэффициент совстречаемости, изменяется от 0 до 1.**

$$S = \frac{a+d}{a+b+c+d}$$

- **коэффициент Жаккара, не учитывает одновременного отсутствия признака при вычислении сходства.**

$$J = \frac{a}{a+b+c}$$

Коэффициенты ассоциативности

амые распространенные:

$$S=\frac{a+d}{a+b+c+d}$$

- **коэффициент совстречаемости, изменяется от 0 до 1.**

- **коэффициент Жаккара,**

$$J=\frac{a}{a+b+c}$$

не учитывает одновременного отсутствия признака при вычислении сходства.

номер признака	X1	X2
1	1	1
2	0	1
3	1	0
4	0	0
5	0	0
6	1	1
7	1	0
8	1	0

	1	0
1	a	b
0	c	d

Коэффициенты ассоциативности

амые распространенные:

$$S = \frac{a+d}{a+b+c+d}$$

- **коэффициент совстречаемости,**

  изменяется от 0 до 1.

- **коэффициент Жаккара,**

$$J = \frac{a}{a+b+c}$$

  не учитывает одновременного

  отсутствия признака при вычислении сходства.

$$S = 0,5$$

$$J = 0,33$$

номер признака	X1	X2
1	1	1
2	0	1
3	1	0
4	0	0
5	0	0
6	1	1
7	1	0
8	1	0

	1	0
1	a	b
0	c	d

Вероятностные коэффициенты сходства

- **По сути сходство не вычисляется, к данным прилагается вероятность. При объединении двух объектов вычисляется информационный выигрыш и те объединения, которые дают мин. выигрыш рассматриваются как один объект.**
- **Недостаток: только для бинарных данных.**

Выбор метода кластеризации

Разработанные кластерные методы образуют 7 основных семейств:

- иерархические агломеративные методы;
- иерархические дивизимные методы;
- итеративные методы группировки;
-методы поиска модальных значений плотности;
- факторные методы;
- методы сгущений;
-методы, основанные на теории графов;
-нейронные сети; деревья решений …

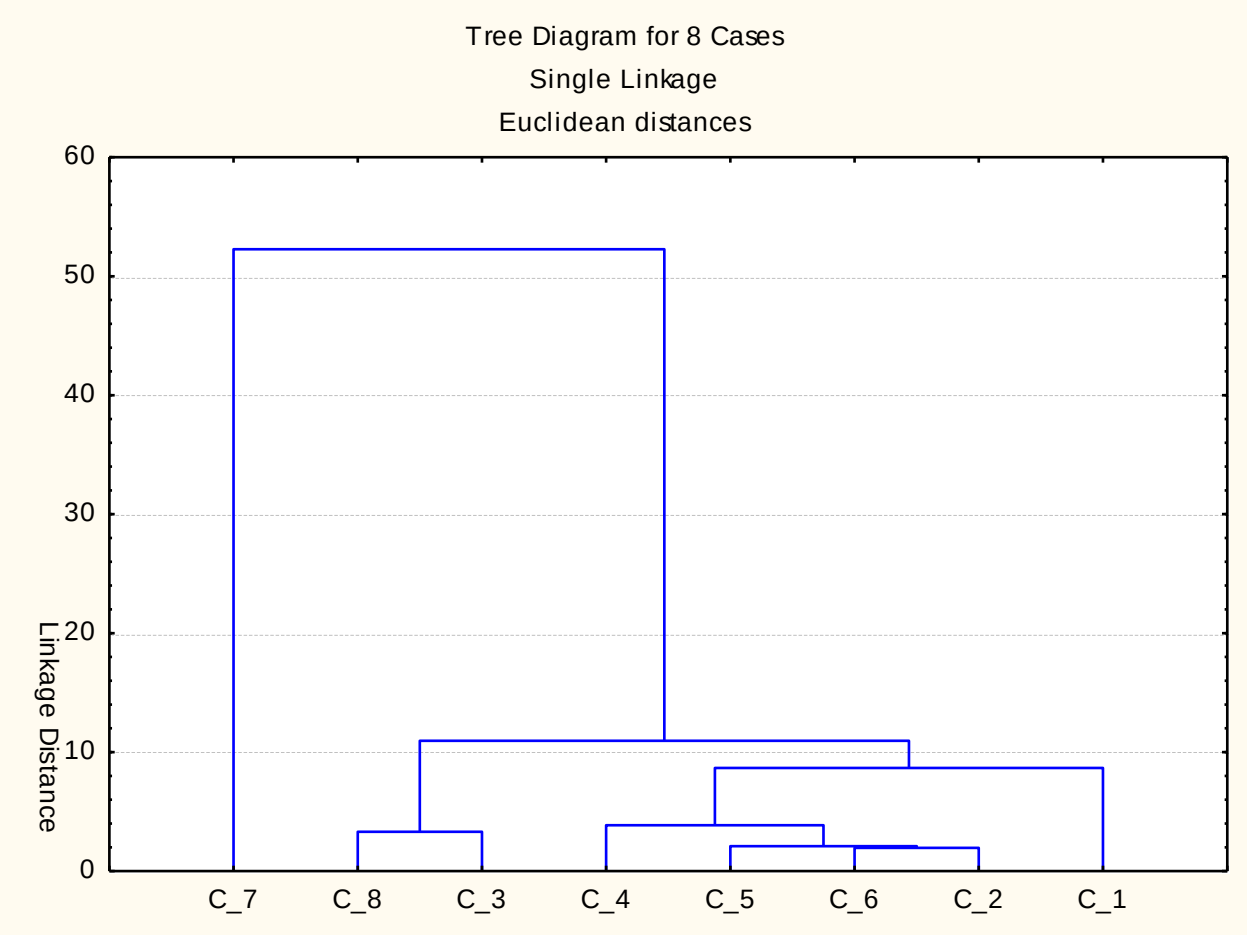
Исходные данные

номер	возраст	стаж	кол-во часов	доход	профессия	пол
1	32	1,6	13,9	9,5	1	2
2	40	0,4	6	8,7	0	1
3	30	2,5	24,6	8,1	1	2
4	36	0,4	6,5	7,8	0	1
5	41	0,3	7,1	10,5	0	2
6	39	0,7	6,7	10,2	0	1
7	24	5	76,2	13,3	2	2
8	30	5	24,1	10,2	1	2

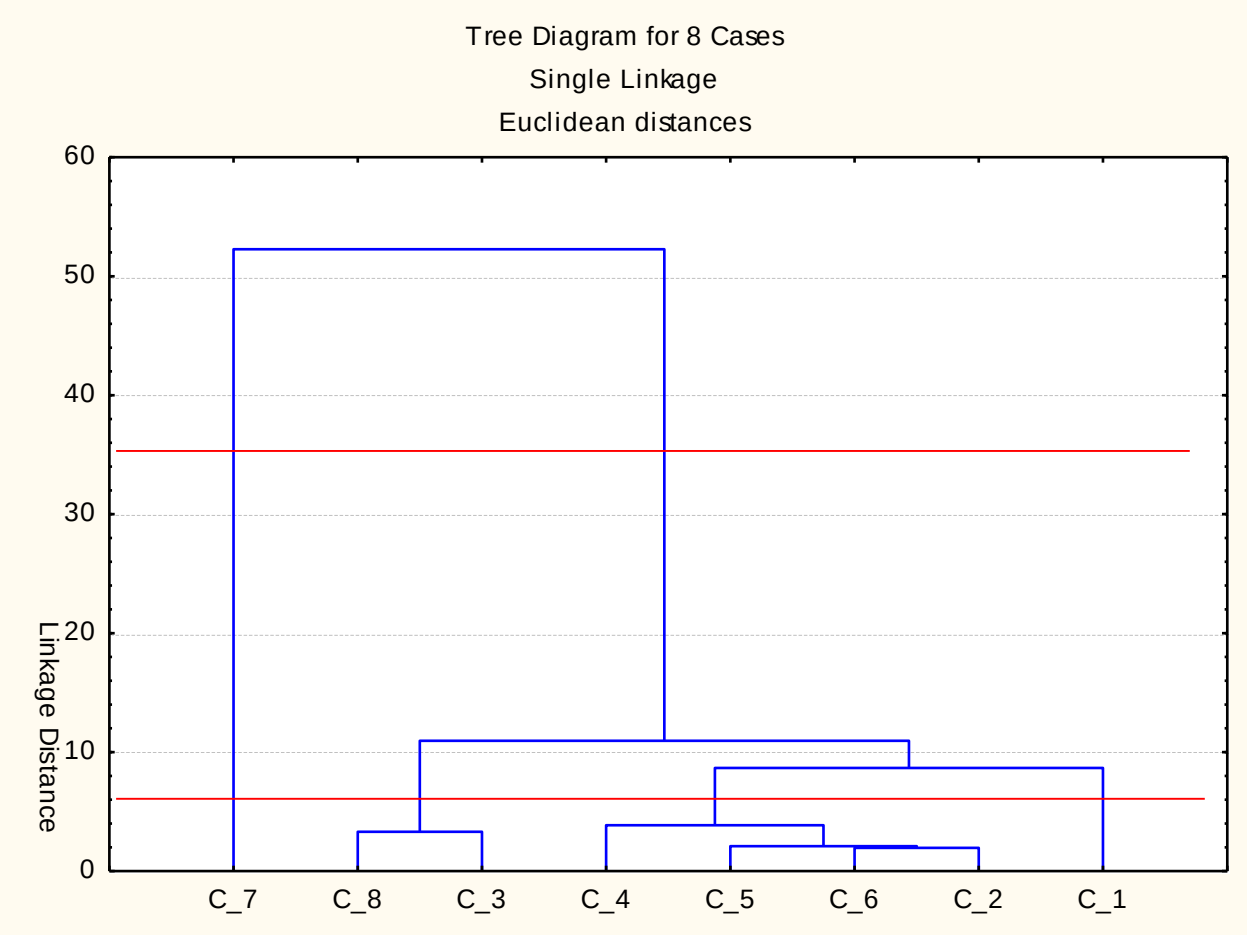
Метод одиночной связи. Матрица расстояний Евклида

	c1	c2	c3	c4	c5	c6	c7	c8
c1	0,0	11,3	11,0	8,7	11,4	10,1	63,0	11,0
c2	11,3	0,0	21,2	4,1	2,3	2,0	72,3	21,2
c3	11,0	21,2	0,0	19,2	20,9	20,2	52,3	3,3
c4	8,7	4,1	19,2	0,0	5,7	3,9	71,1	19,3
c5	11,4	2,3	20,9	5,7	0,0	2,1	71,4	20,8
c6	10,1	2,0	20,2	3,9	2,1	0,0	71,3	20,1
c7	63,0	72,3	52,3	71,1	71,4	71,3	0,0	52,5
c8	11,0	21,2	3,3	19,3	20,8	20,1	52,5	0,0

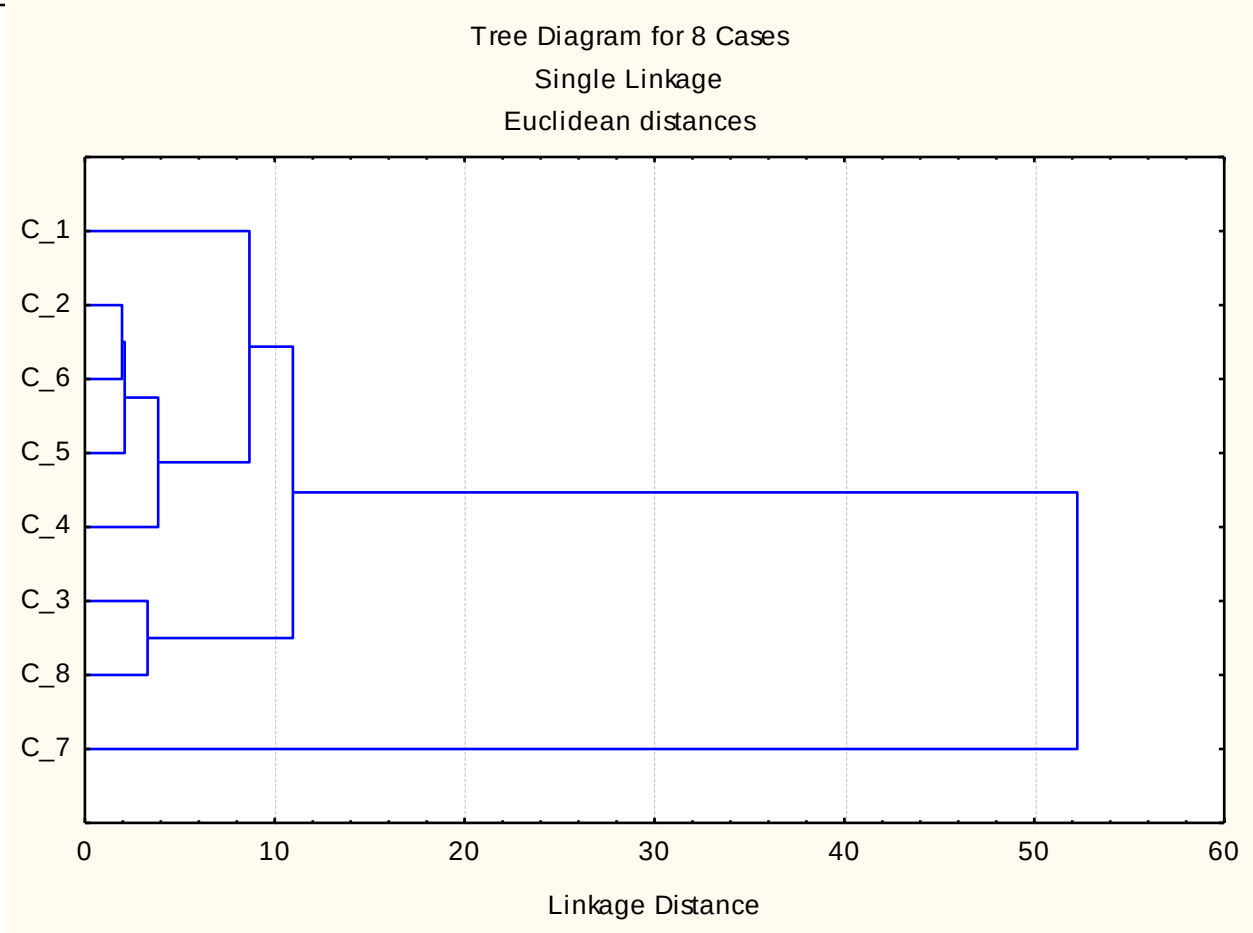
Метод одиночной связи. Дендрограмма



Метод одиночной связи. Дендрограмма



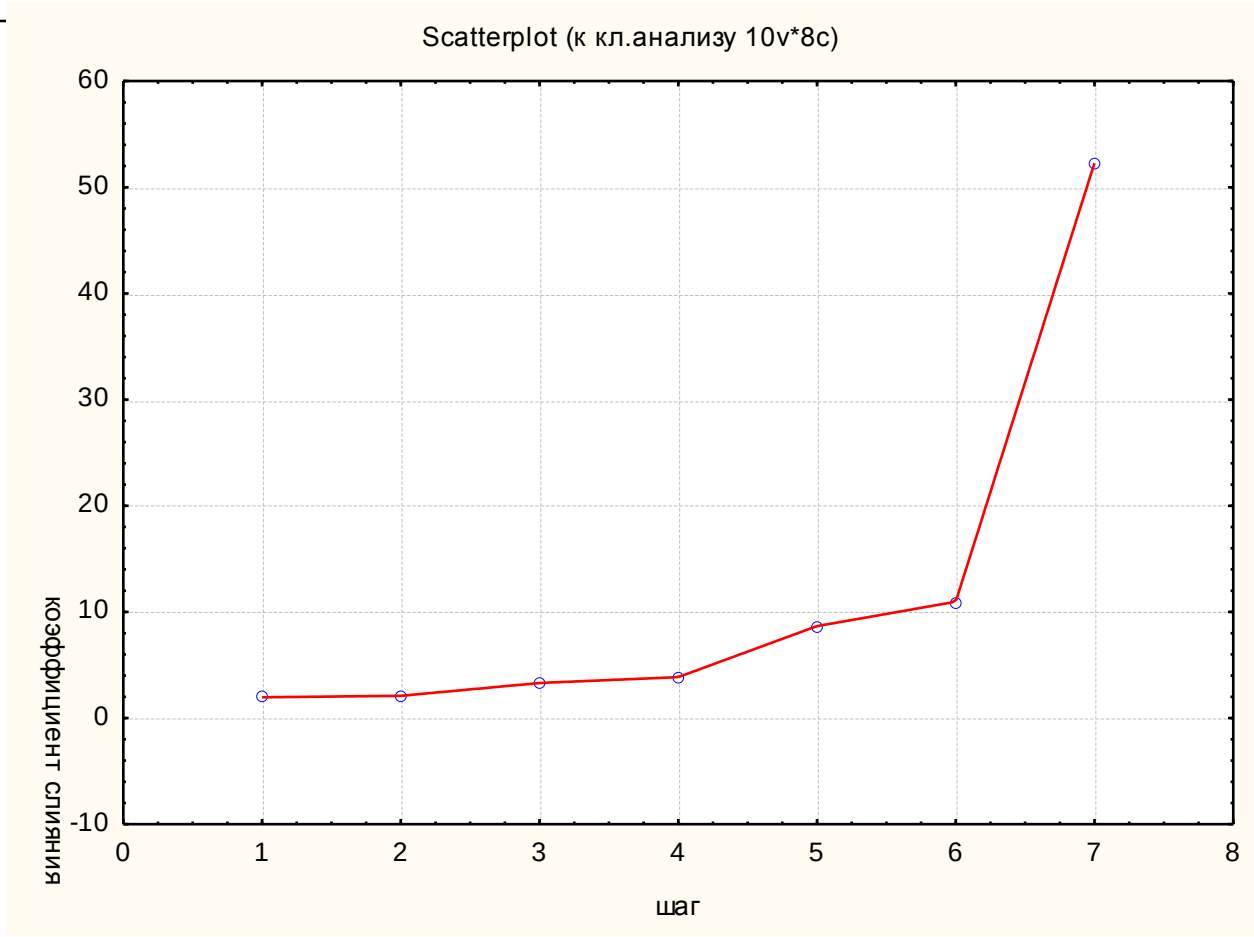
Метод одиночной связи. Дендрограмма



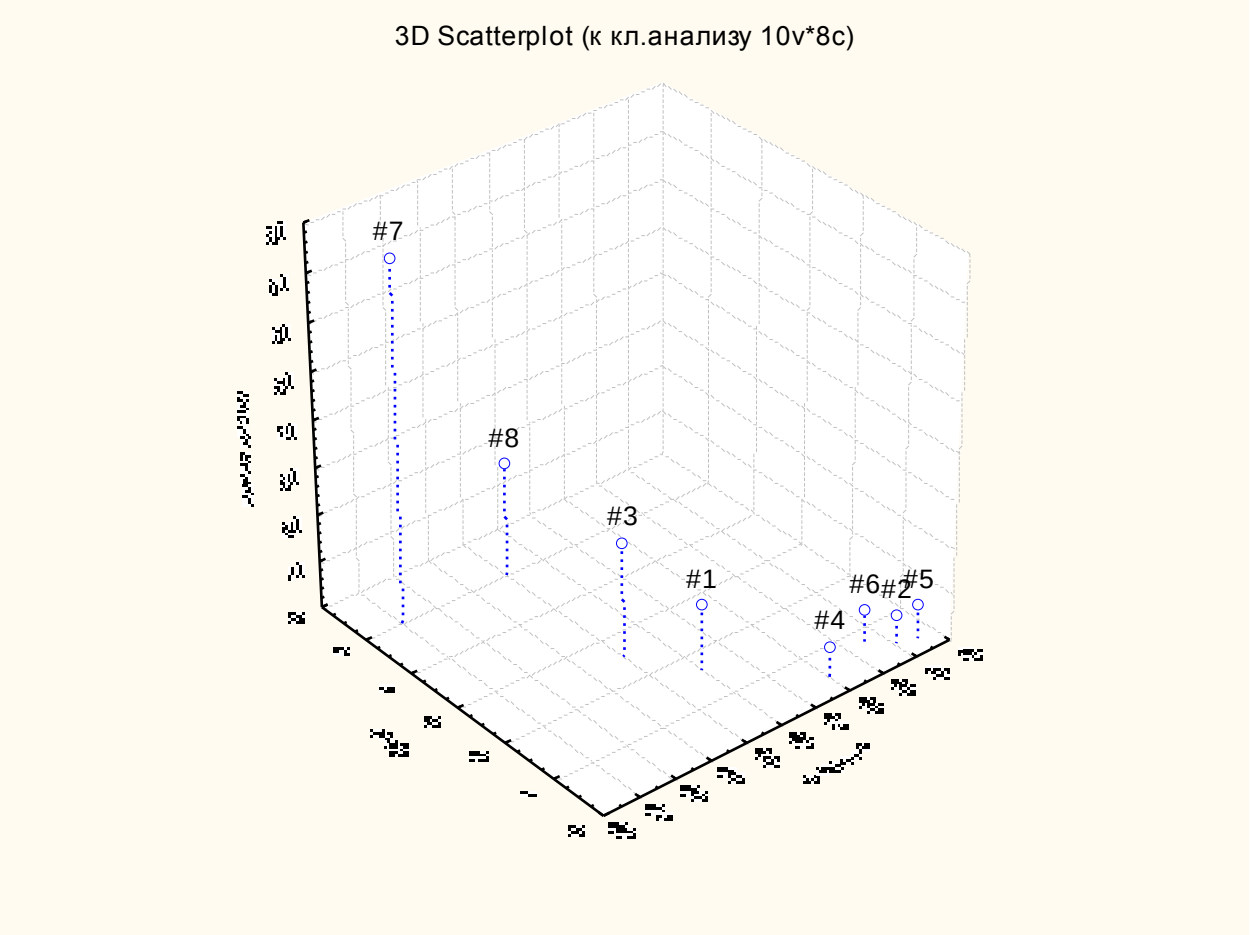
Метод одиночной связи. Схема объединения

Amalgamation Schedule (к кл.анализу) Single Linkage Euclidean distances								
	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.
1,95703	C_2	C_6						
2,10000	C_2	C_6	C_5					
3,30302	C_3	C_8						
3,85875	C_2	C_6	C_5	C_4				
8,66544	C_1	C_2	C_6	C_5	C_4			
10,9585	C_1	C_2	C_6	C_5	C_4	C_3	C_8	
52,2671	C_1	C_2	C_6	C_5	C_4	C_3	C_8	C_7

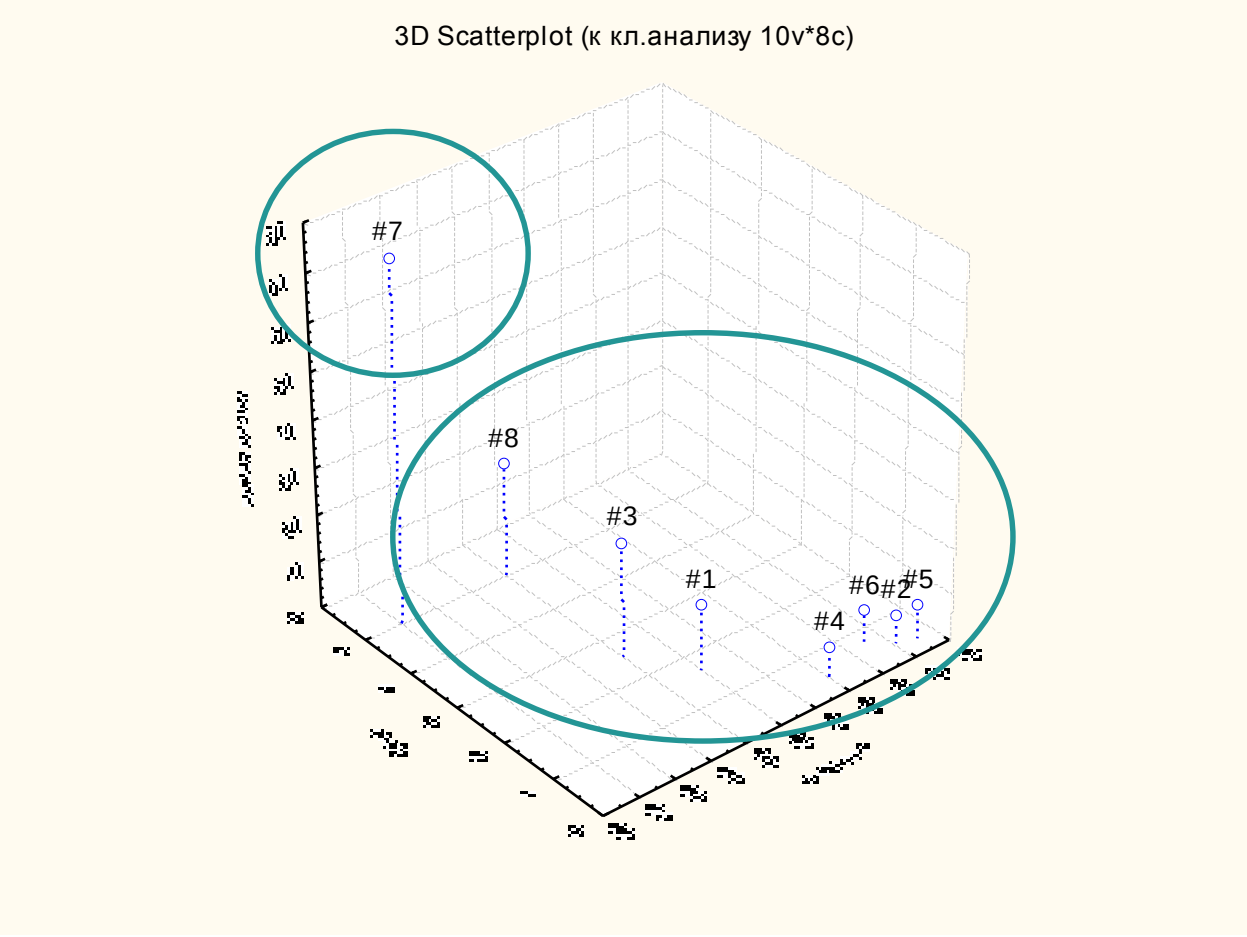
Метод одиночной связи. График связи



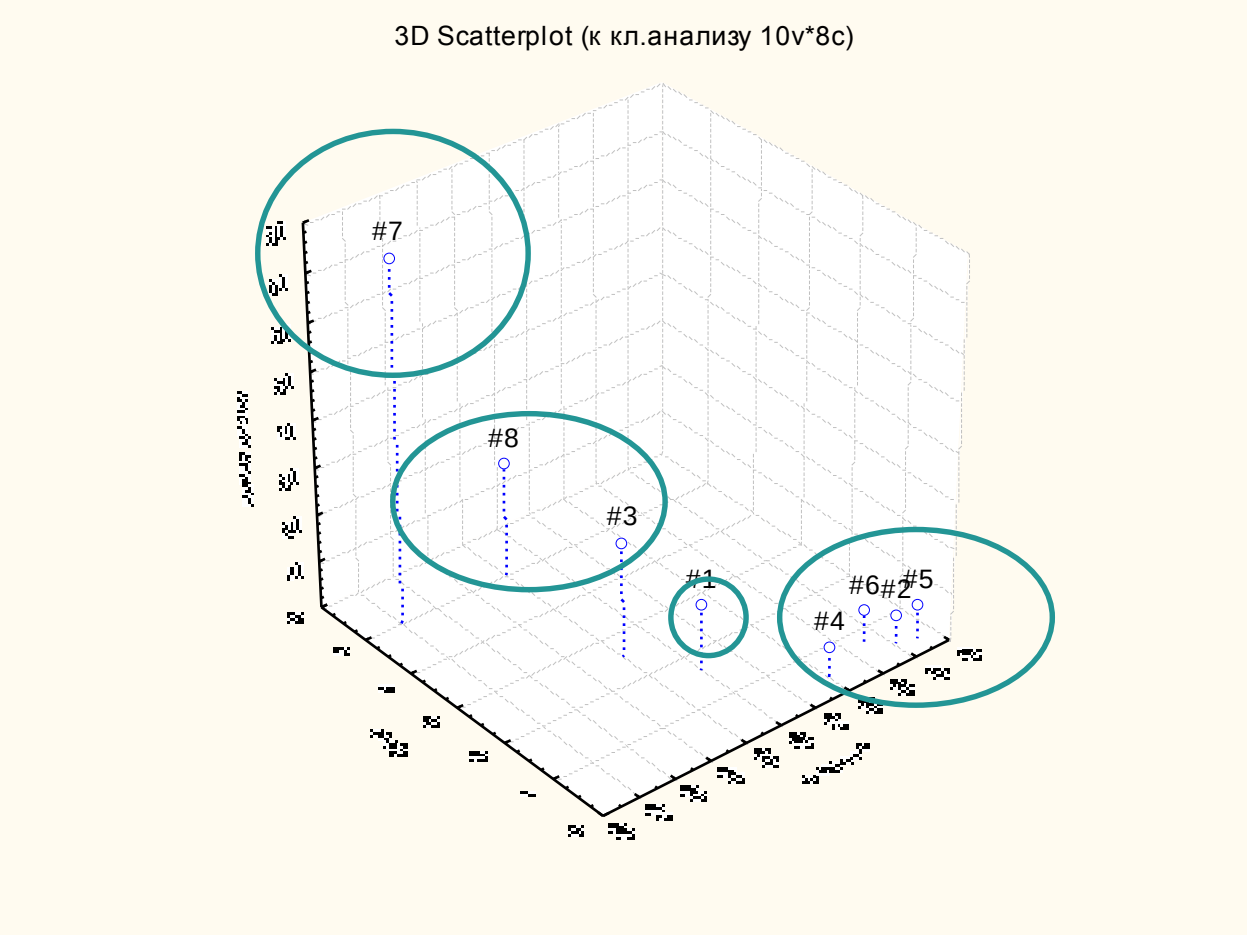
Метод одиночной связи. Диаграмма рассеяния



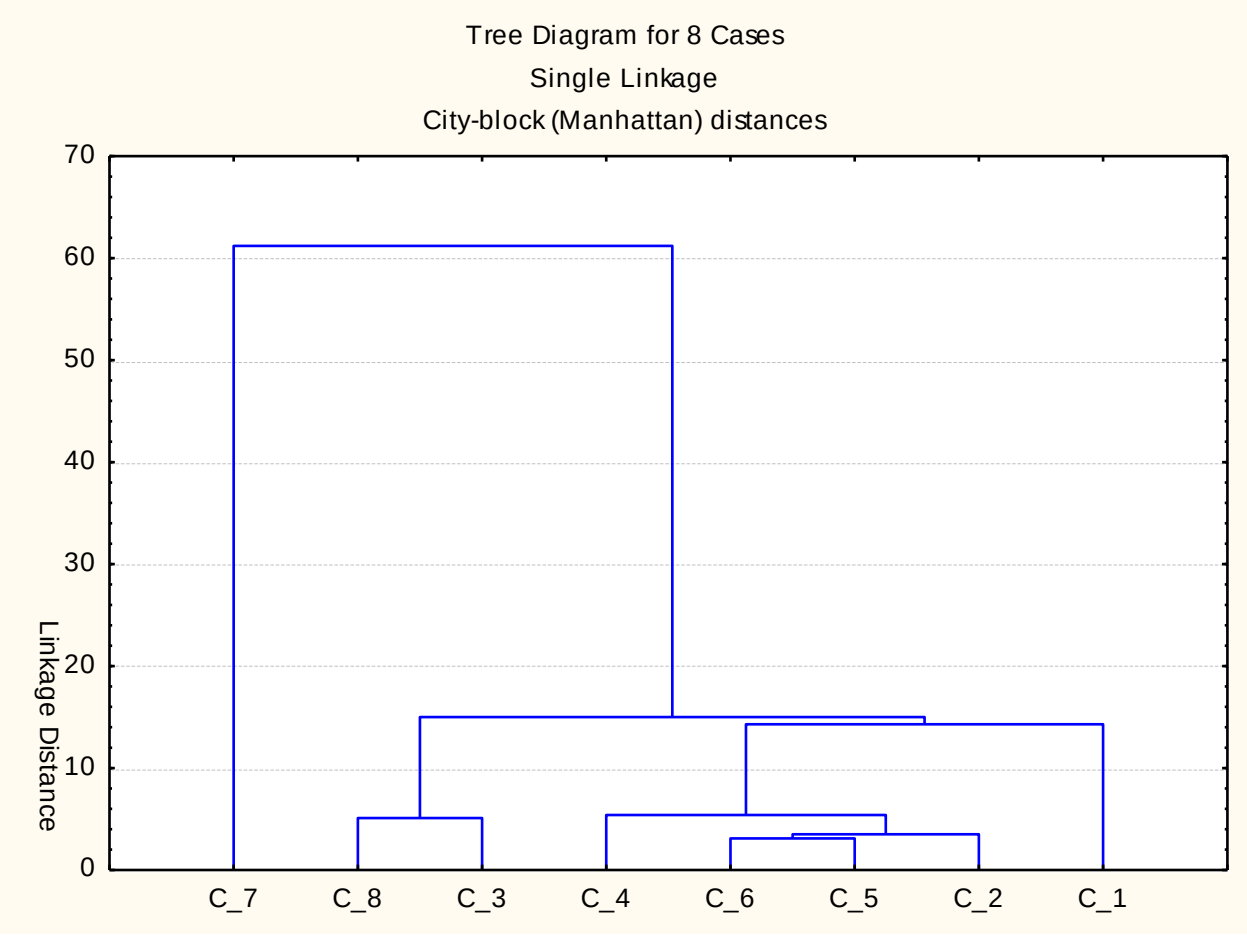
Метод одиночной связи. Диаграмма рассеяния



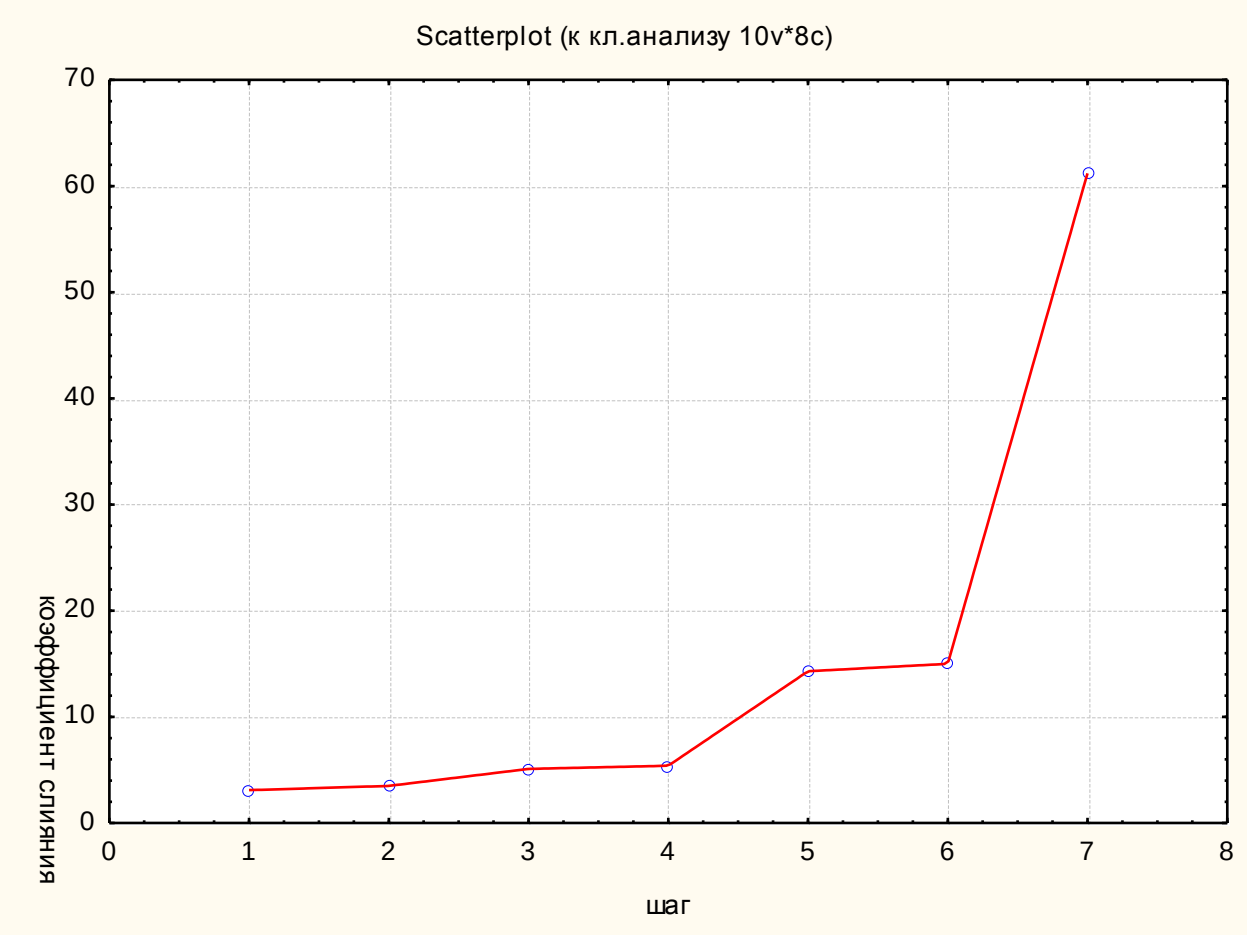
Метод одиночной связи. Диаграмма рассеяния



Метод одиночной связи. Дендрограмма



Метод одиночной связи. График слияния



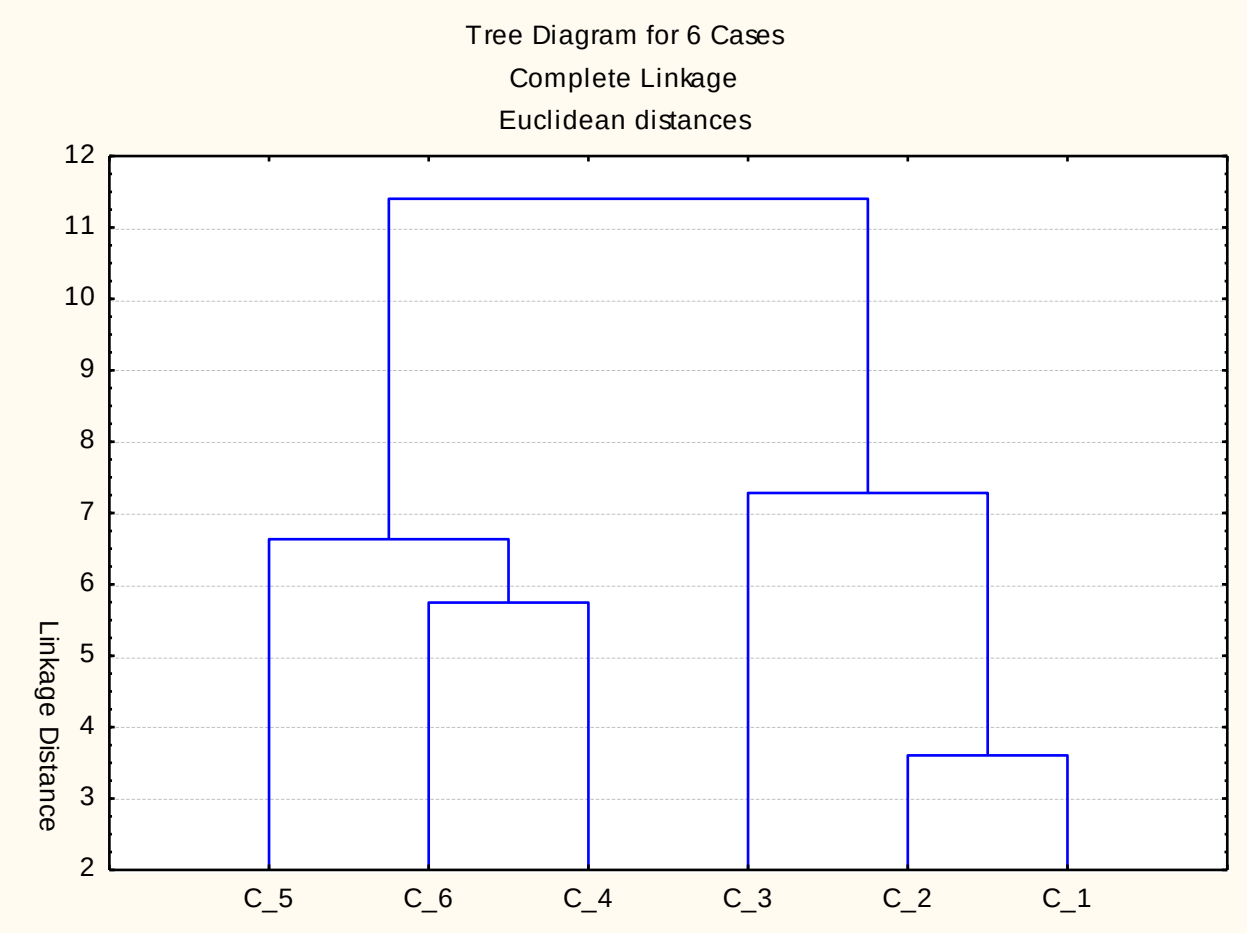
Метод полной связи. Исходные данные

	признак 1	признак 2	признак 3	признак 4
1	9	21	45	55
2	12	23	45	55
3	13	21	39	56
4	9	25	46	49
5	6	27	41	50
6	11	28	44	53

Метод полной связи. Матрица расстояний Евклида

Euclidean distances (к кл.анализу)						
	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0,00	3,61	7,28	7,3	9,3	7,62
C_2	3,61	0,00	6,5	7,1	9,6	5,57
C_3	7,28	6,48	0,0	11,4	11,2	9,33
C_4	7,28	7,07	11,4	0,0	6,2	5,74
C_5	9,27	9,64	11,2	6,2	0,0	6,63
C_6	7,62	5,57	9,3	5,7	6,6	0,00

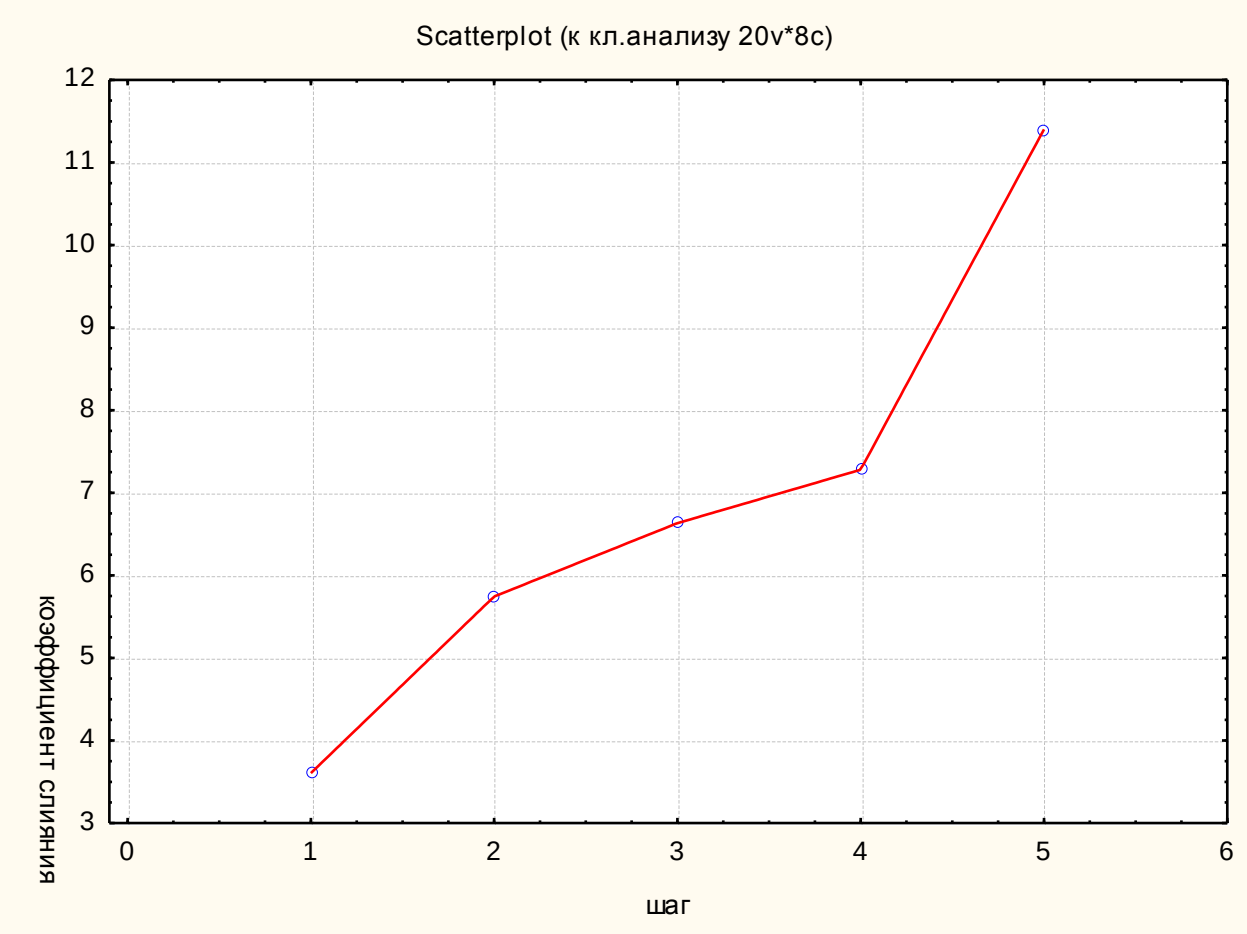
Метод полной связи. Дендрограмма



Метод полной связи. Схема объединения

Amalgamation Schedule (к кл.анализу) Complete Linkage Euclidean distances						
	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.	Obj. No.
3,60555	C_1	C_2				
5,74456	C_4	C_6				
6,63325	C_4	C_6	C_5			
7,28011	C_1	C_2	C_3			
11,4017	C_1	C_2	C_3	C_4	C_6	C_5

Метод полной связи. График объединения



Метод Варда (Уорда)

-Метод использует методы дисперсионного анализа для оценки расстояний между кластерами.

-Метод минимизирует сумму квадратов отклонений для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге.

Метод Варда (Уорда)

Сумма квадратов отклонений:

$$\sigma = \sum_{k=1}^{p} \{x_{kj} - (\frac{1}{n} \sum_{j=1}^{n} x_{kj})\}^2$$

x_{kj} - значение k-го признака j-го объекта.

n – число объектов в кластере;

p – число признаков.

Метод Варда (Уорда)

- Метод эффективен, однако он стремится создавать кластеры малого размера.
- В результате использования метода примерно равные кластеры по размеру и имеющие гиперсферическую форму.

Метод К-средних

Постановка задачи:

Объекты $X_1, X_2, \ldots, X_n$ требуется разбить на заданное число К (K<n) однородных в смысле некоторой метрики классов.

0 шаг. Выбирается К эталонных точек (центров групп). Объекты в качестве эталонных точек выбираются случайным образом.

$$E^{(0)} = \{e_1^0, e_2^0, ..., e_k^0\} \qquad e_i^0 = X_i$$

Каждой эталонной точке приписывается вес:

$$w_i^{(0)} = 1, i = 1,2,..., k$$

1 шаг-v. Извлекается случайным образом объект X_{k+1}

и выясняется к какой из эталонных точек он ближе всего. Самая близкая к X_{k+1} эталонная точка заменяется центром тяжести старого эталона и объекта X_{k+1} .

(Центр тяжести вычисляется как среднее признаков, измеренных на объектах). Вес этой эталонной точки возрастает на 1, все остальные эталонные точки остаются без изменения.

Правило вычисления эталона на v итерации:

$$e_i^{(v)} = \begin{cases} \dfrac{w_i^{(v)} e_i^{(v-1)} + X_{k+v}}{w_i^{(v)} + 1} \\ e_i^{(v-1)} \end{cases}$$

-для самой близкой к X_{k+1} эталонной точки, для всех остальных эталонных точек.

Если объект одинаково близок к нескольким эталонным точкам, то устанавливается правило его отнесения, например, к первой по порядковому номеру.

При достаточно большом числе итераций, пересчет дальнейший эталонных точек практически не приводит к их изменению, т.е. имеет место сходимость эталонных точек к некоторому пределу при $\nu \to \infty$.

Пример для метода К-средних

	1	2	3	4	5	6
P1	10	15	5	25	20	25
P2	30	35	20	30	25	20
P3	100	105	115	120	110	115
P4	70	75	65	60	80	65

Достоинства итеративных методов:

- Кластеры одного ранга, не являются вложенными и частью иерархии;
- не допускается перекрытие кластеров;
- можно классифицировать большой объем наблюдений.

Оценка полученных результатов

Методы проверки обоснованности кластерных решений

кофенетическая корреляция;

- тесты значимости для признаков, используемых при создании кластеров;
- повторная выборка;
- тесты значимости для внешних признаков;
- методы Монте-Карло.

Ограничения использования кластерного анализа данных

- Методы и процедуры кластерного анализа, как правило не имеют достаточного статистического обоснования.

- Разные кластерные методы могут порождать различные решения для одних и тех же данных.

- Цель кластерного анализа – поиск существующих структур в данных. В то же время КА привносит структуру в анализируемые данные. Поэтому необходимо отличить «реальные» группировки от навязанных методом кластеризации.