Кластерный анализ

Наумов Д.А., доц. каф. КТ

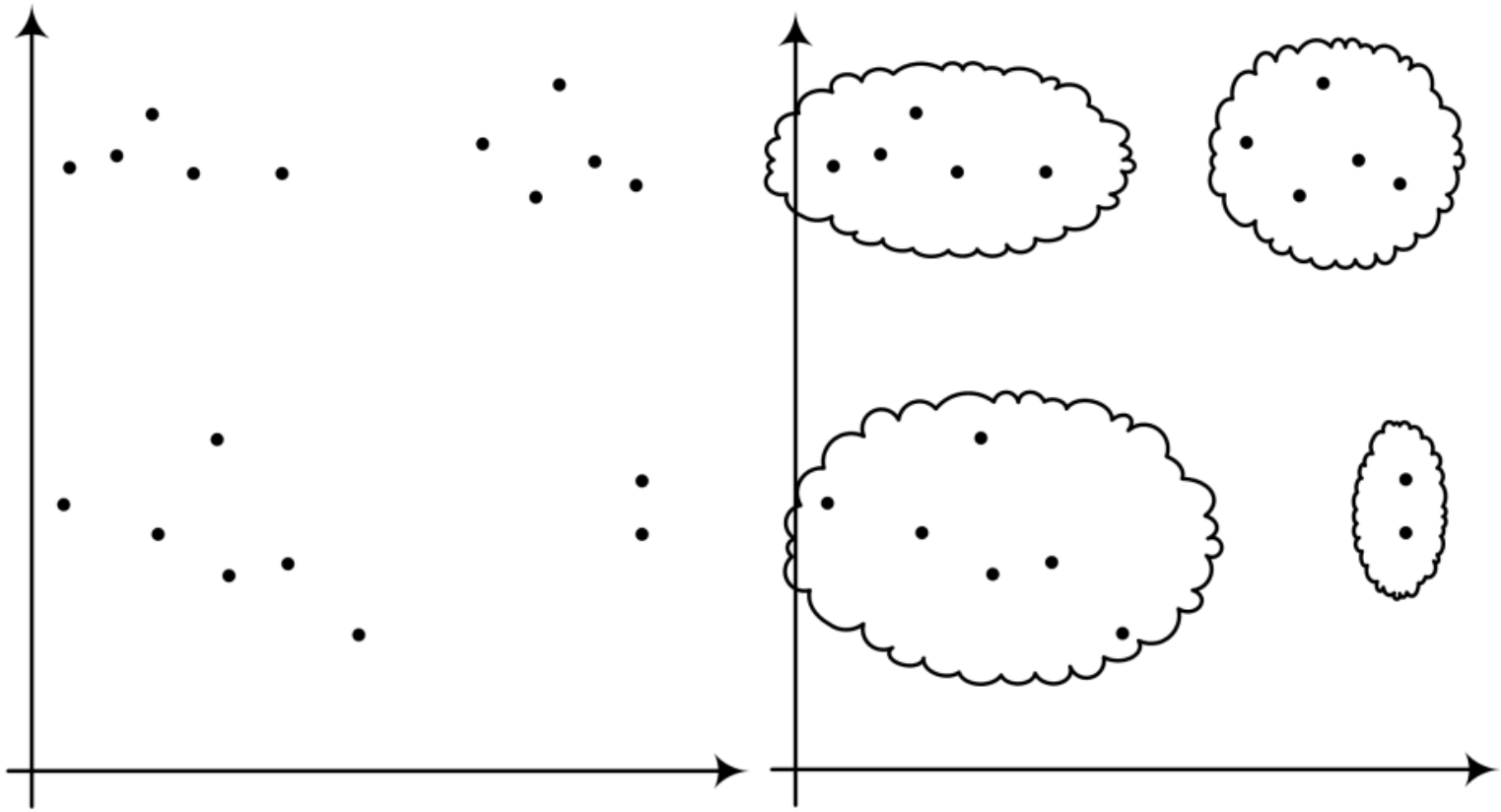
Экспертные системы и искусственный интеллект, 2019

Содержание лекции

Кластеризация

Задача кластеризации

частично сгруппированные точки на плоскости или в пространстве большей размерности разбиваются на близкорасположенные группы.



Кластеризация

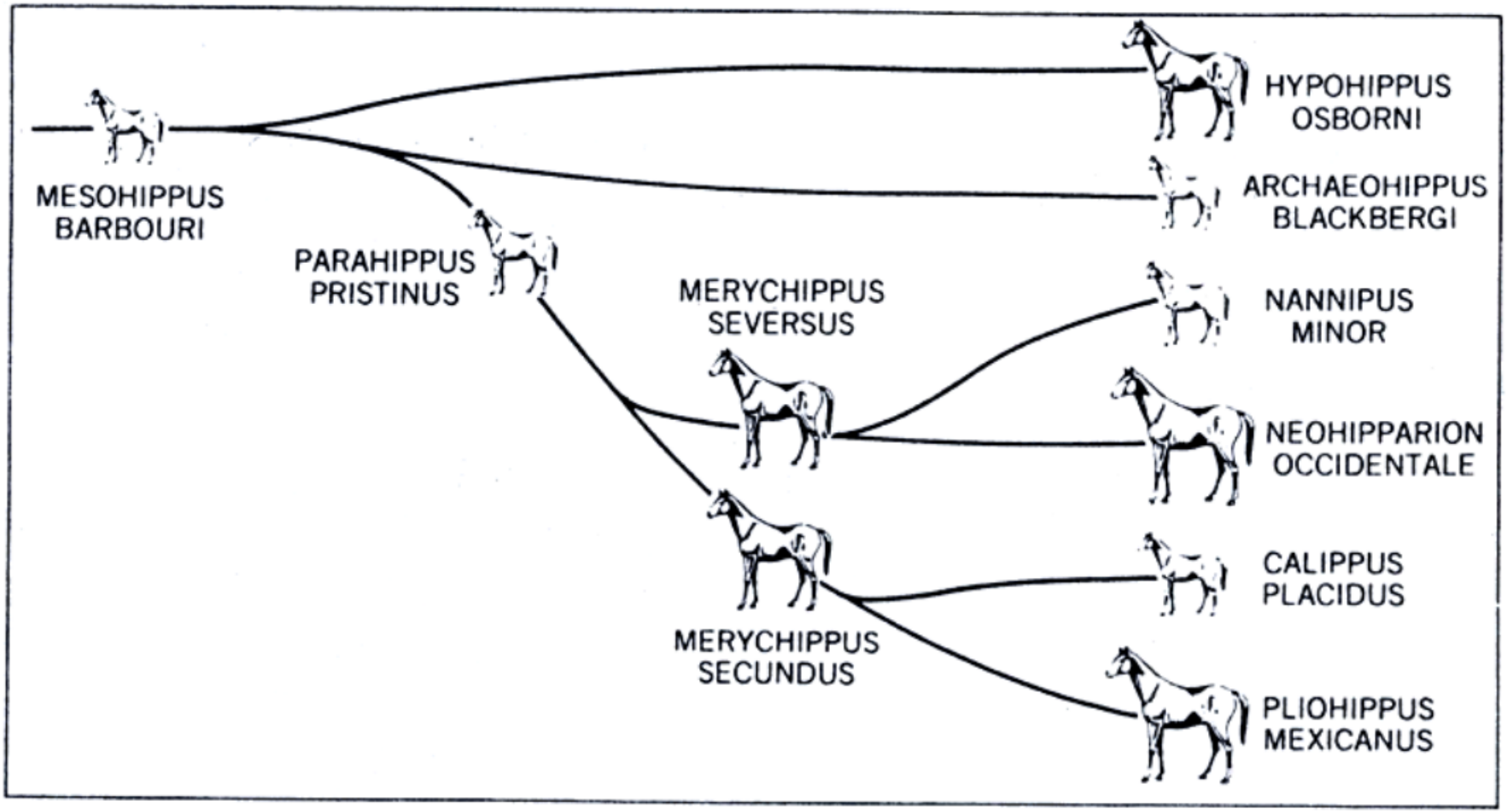


Figure 12 An example of a taxonomic tree (from Sokal, 1966).

Задачи кластеризации

- лежат в основе анализа данных (data mining);
  - анализ покупок в супермаркетах;
  - разделение документов на жанры и тематики;
  - контекстная рекалама и анализ предпочтений пользователя.
- распознование образов: получение границ областей и выделение общих принаков;
- биоинформатика: анализ длинных последовательностей атомов в белках
  - разбиение генов на кластеры;
  - выделение генов на группы по функциональной схожести;
  - выделение генов, отвечающие за биологическое свойство.
- маркетологические исследования для сегментации рынка;
- анализ социальныз сетей;

Кластеризация

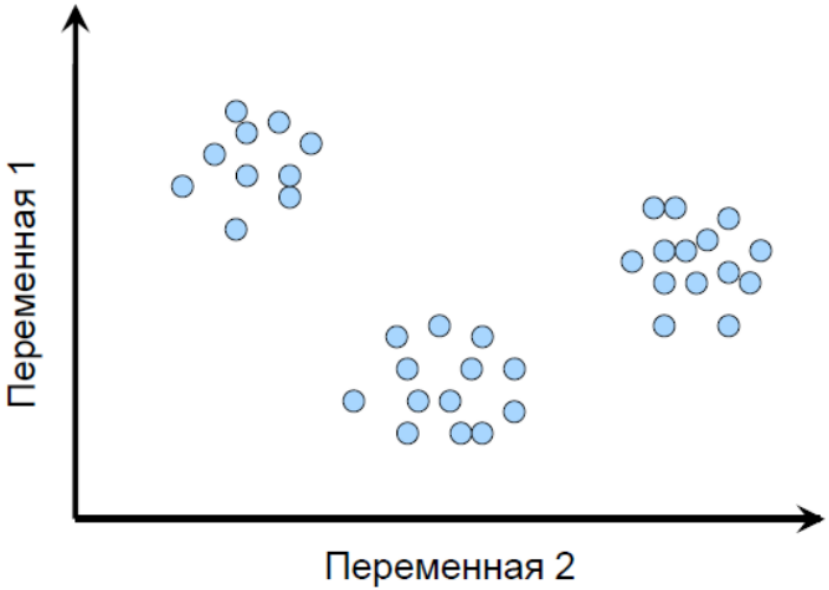
Кластеризация - типичная задача статистического анализа

задача классификации объектов одной природы в несколько групп так, чтобы объекта одной группы обладали одним и тем же свойством. Под свойством понимается близоть друг к другу относительно выбранной метрики.

Сложность задачи кластеризации:

- высокая размерность реальных задач (сотни и тысячи);
- необхождим компромис: число точек и размерность пространства.

Кластерный анализ в теории



Кластерный анализ на практике

Кластеризация

Пусть дан набор тестовых примеров

$$X = \{x_1, \ldots, x_n\}$$

и функция расстояния между ними

$$\rho : X \times X \to \mathbb{R}.$$

Требуется разбить X на непересекающиеся подмножества, которые, собственно, и называются кластерами, так, чтобы каждое подмножество состояло из *близких объектов*, а объекты разных подмножеств *существенно различались*.

- что такое "близкие объекты"?
- что такое "существенно различающиеся объекты"?

Меры близости между объектами

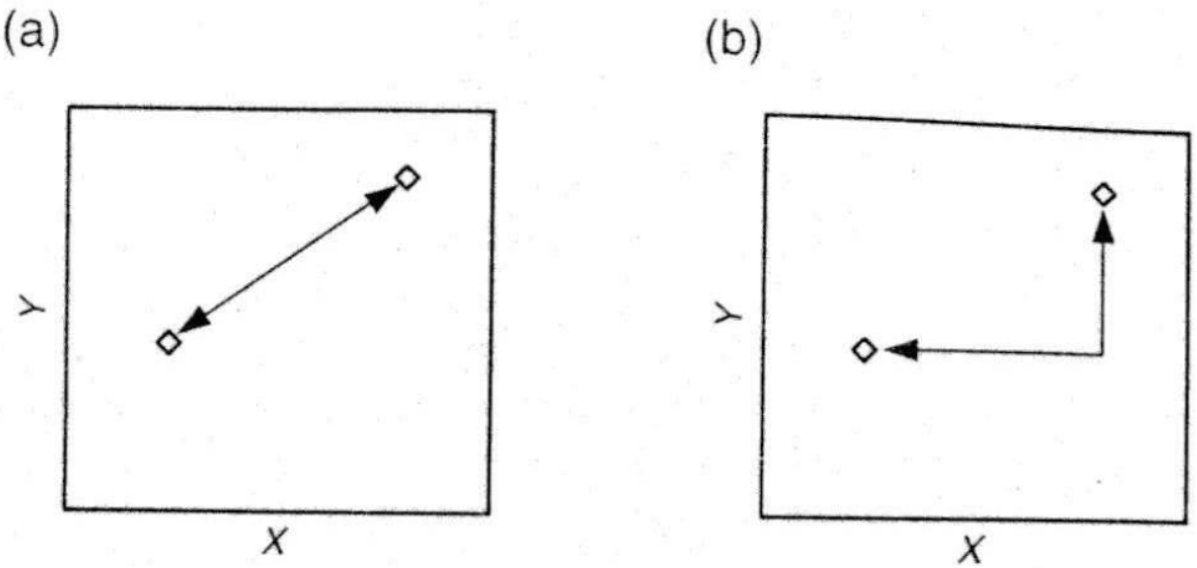


Figure 15.2 Different measures of distance between two points: (a) Euclidean distance, (b) Manhattan or city block distance

Меры близости между объектами

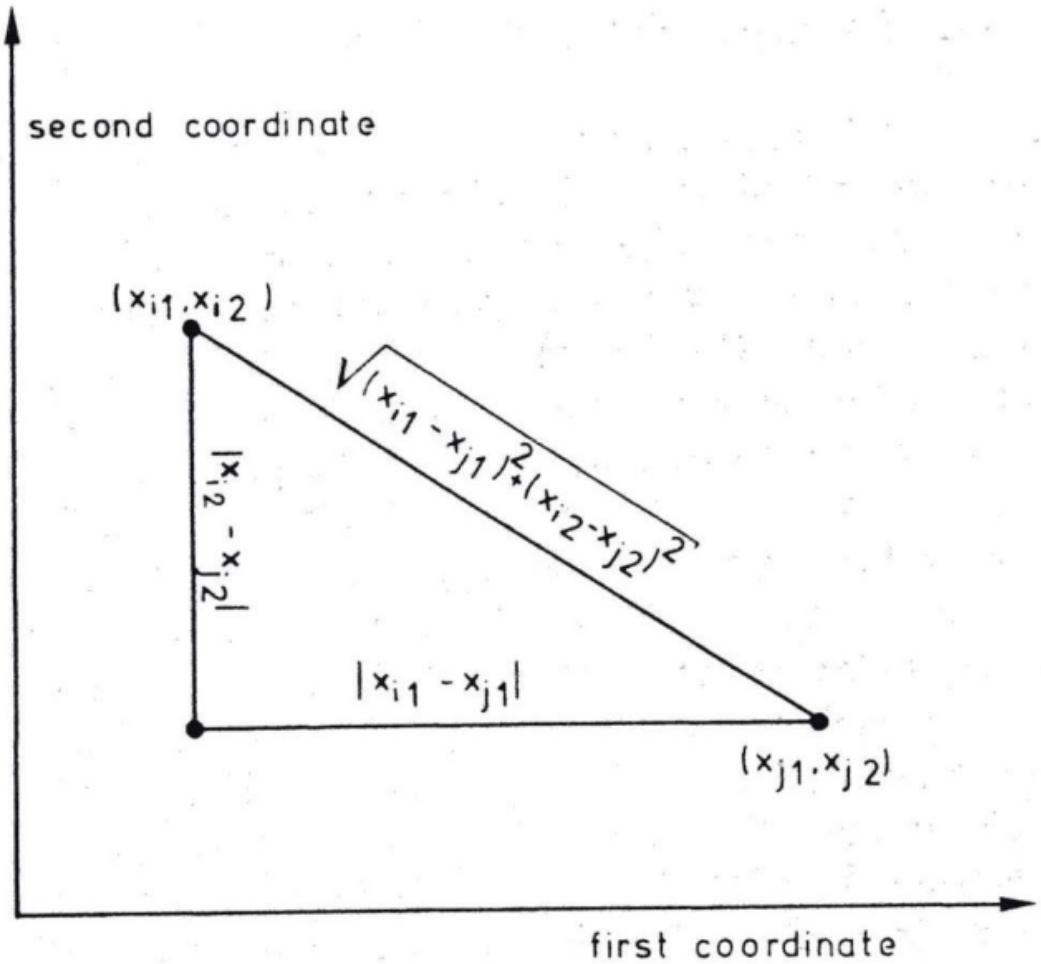


Figure 6 Illustration of the Euclidean distance formula.

Меры близости между объектами

Меры близости между объектами (меры подобия)

Показатели	*Формулы*
Для количественных шкал	
Линейное расстояние	$d_{1ij} = \sum_{l=1}^{m} \lvert x_i^l - x_j^l \rvert$
Евклидово расстояние	$d_{Eij} = \left(\sum_{l=1}^{m} (x_i^l - x_j^l)^2 \right)^{1/2}$
Квадрат евклидово расстояние	$d_{Eij}^2 = \sum_{l=1}^{m} (x_i^l - x_j^l)^2$
Обобщенное степенное расстояние Минковского	$d_{Pij} = \left(\sum_{l=1}^{m} (x_i^l - x_j^l)^p \right)^{1/p}$
Расстояние Чебышева	$d_{ij} = \max_{1 \le i,j \le l} \lvert x_i - x_j \rvert$
Расстояние городских кварталов (Манхэттенское расстояние)	$d_M(x_i, x_j) = \sum_{l=1}^{n} \lvert x_i^l - x_j^l \rvert$

Расстояние tf-idf

tf-idf

мера близости двух текстовых документов

Документ представляется в виде вектора из n термов с некоторыми весами. Разные подходы к анализу текстов различаются в том:

- что такое терм;
- как определять веса.

Термы

(обычно) слова, встречающиеся в документе.

Документ превращается при этом в неупорядоченный набор слов (bag of words).

Сложные структуры в качестве термов? нецелесообразно!

- нужно будет слишком много текстов,
- результат не будет существенно лучше bag-of-words подхода.

Расстояние tf-idf

Для определения весов обычно используют два основных подхода:

- либо бинарный атрибут со значениями 01 (есть слово или нет слова),
- либо весовую функцию, меру tdf (term frequency inverse document frequency).

Мера tf-idf

- была предложена в начале 1970-х годов и с тех пор активно используется в анализе текстовой информации и information retrieval;
- состоит из двух других: tf (частота терма, term frequency) и idf (обратная частота терма в документах, inverse document frequency).

Расстояние tf-idf

Частота терма

доля числа появлений этого терма по отношению к размеру всего документа.

$$\mathrm{tf}(t_k, d_j) = \frac{\#(t_k, d_j)}{\sum_k \#(t_k, d_j)},$$

где #(tk,dj) - число, показывающее, сколько раз терм tk встречается в документе dj.

Расстояние tf-idf

Обратная частота терма

показывает, насколько терм вообще важен, насколько он характерен для данного массива текстов.

Чем реже встречается терм в имеющемся массиве, тем он характернее.

$$\mathrm{idf}(t_k, d_j) = \log \frac{|D|}{\#_D t_k},$$

где где D - имеющийся набор данных, а #Dtk - количество документов из D, в которых хотя бы однажды встречается tk.

Расстояние tf-idf

Мера tdf для терма tk и документа dj в массиве D равна:

$$\mathrm{tfidf}(t_k,d_j) = \mathrm{tf}(t_k,d_j)\mathrm{idf}(t_k,d_j) = \frac{\#(t_k,d_j)}{\sum_k \#(t_k,d_j)} \log \frac{|D|}{\#_D t_k}$$

Вектор весом можно нормализовать:

$$w_{kj} = \frac{\mathrm{tfidf}(t_k,d_j)}{\sqrt{\sum_{s=1}^{r}(\mathrm{tfidf}(t_k,d_j))^2}}$$

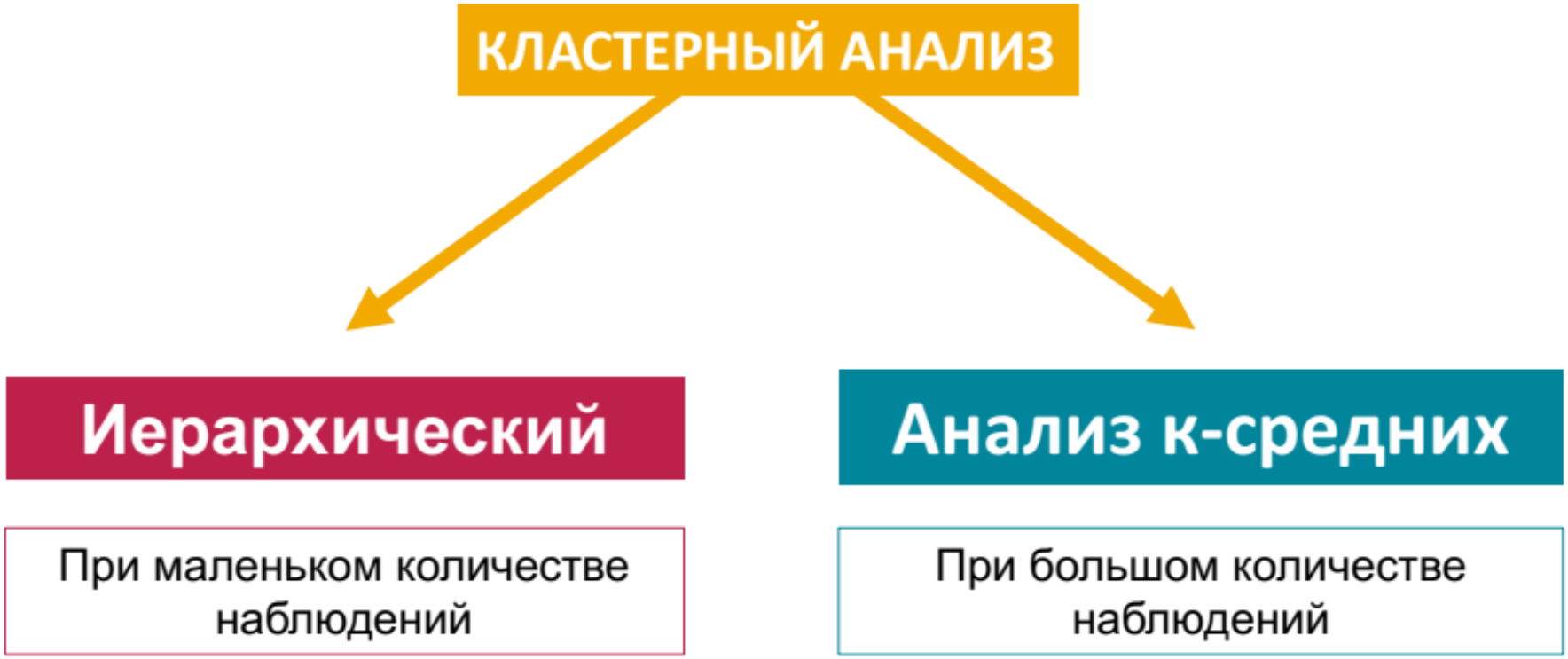
Теперь документ можно представить в виде вектора, размерность которого равна количеству термов (важно сохранять словарь не слишком большим за счет удачного выбора термов).

Расстояние между документами

Расстояние между документами - используется не простое декартово расстояние, а угол между векторами, косинусоидальная мера похожести (cosine similarity measure):

$$\theta(d_1, d_2) = \arccos \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$$

Кластерный анализ



- иерархическая кластеризация - алгоритм последовательно строит кластеры из уже найденных кластеров;
- неиерархическая кластеризация - алгоритм пытается распознать всю структуру сразу или строит кластеры один за другим, не разделяя и не соединяя уже выделенные кластеры.

Кластерный анализ

Иерархическая кластеризация:

- агломеративная: алгоритм начинает с индивидуальных элементов (кластеров из одного элемента), а затем последовательно объединяет их, получая требуемую структуру;
- разделительная, когда алгоритм начинает с одного кластера, содержащего все точки, а потом последовательно делит его на части.

Неиерархические методы, как правило, стремятся оптимизировать некую целевую функцию, которая описывает качество кластеризации.

- агоритмы, основанные на методах теории графов;
- алгоритм EM;
- алгоритм k-средних;
- нечеткие алгоритмы.

Иерархическая кластеризация. Пусть нужно кластеризовать точки x1,x2, . . . ,xn в некотором метрическом пространстве с метрикой.

1. на первом шаге мы считаем каждую точку отдельным кластером;
2. ближайшие точки объединяем и далее относимся к ним как к единому кластеру.
3. при итерации этого процесса получается дерево, в листьях которого отдельные точки, а в корне кластер, содержащий все точки вообще.

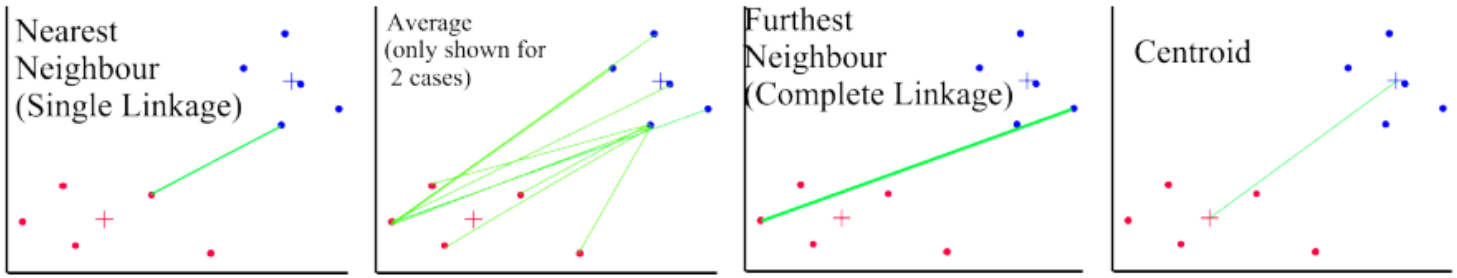
HierarchyCluster($X = \{x_1, \ldots, x_n\}$):

1. Инициализируем $C = X$, $G = X$.
2. Пока в C больше одного элемента:
   а) Выбираем два элемента C c_1 и c_2, расстояние между которыми минимально.
   б) Добавляем в G вершину c_1c_2, соединяем её с вершинами c_1 и c_2.
   в) $C := C \cup \{c_1c_2\} \setminus \{c_1, c_2\}$.
3. Выдаём G.

Расстояние между кластерами

Метод кластеризации – это способ вычисления расстояний между кластерами. Существуют следующие основные методы кластеризации:

- Межгрупповая связь (Between-groups linkage)
- Внутригрупповая связь (Within-groups linkage)
- Ближайший сосед (Nearest neighbor)
- Самый дальний сосед (Furthest neighbor)
- Центроидная кластеризация (Centroid clustering)
- Медианная кластеризация (Median clustering)
- Метод Варда (Уорда)(Ward's method)



Расстояние между кластерами

Стандартизация данных

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_{X_j}}$$

Z-стандартизация	Из значений вычитается среднее и затем они делятся на стандартное отклонение.
Разброс от -1 до 1	Линейным преобразованием переменных добиваются разброса значений от -1 до 1.
Разброс от 0 до 1	Линейным преобразованием переменных добиваются разброса значений от 0 до 1.
Максимум 1	Значения переменных делятся на их максимум.
Среднее 1	Значения переменных делятся на их среднее.
Стандартное отклонение 1	Значения переменных делятся на стандартное отклонение.

Расстояние между кластерами

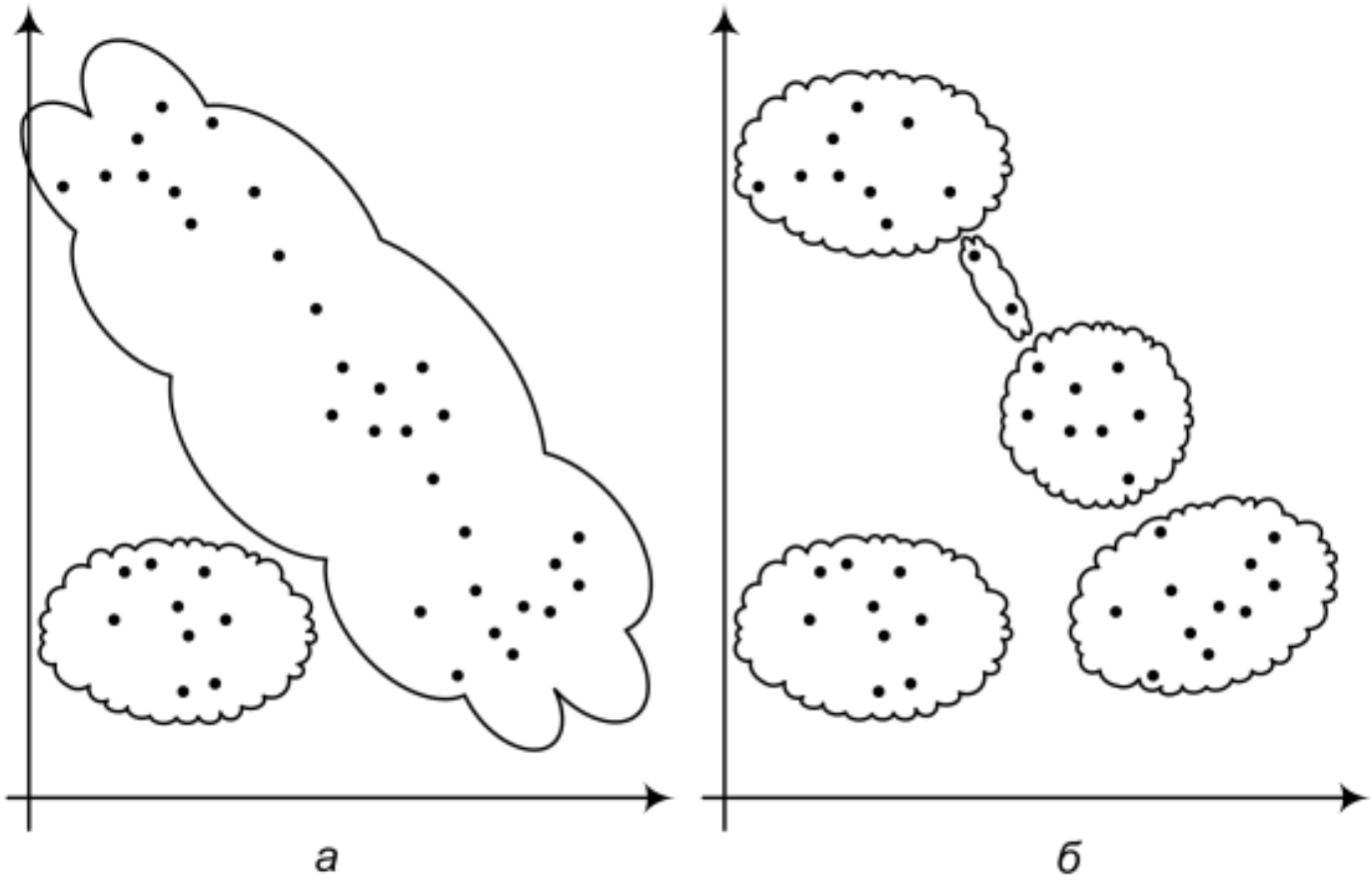


Рис. 6.3. Виды иерархической кластеризации:
а — single-link; *б* — complete-link

Алгоритм кластеризации на основе теории графов

Определение. *Остовное дерево* (spanning tree) графа $G = \langle V, E \rangle$ — это связный подграф G, содержащий все вершины G и являющийся деревом. *Минимальное остовное дерево* (minimal spanning tree) графа $G = \langle V, E \rangle$ с заданными на рёбрах весами $w : E \to \mathbb{R}$ — это такое остовное дерево T, что его суммарный вес не превосходит суммарного веса любого другого остовного дерева T':

$$\forall T' \quad \sum_{e \in T'} w(e) \geqslant \sum_{e \in T} w(e).$$

1. построить минимальное остовное дерево;
2. выкидывать из него ребра максимального веса до тех пор, пока не получится нужное число кластеров.

Сколько ребер выбросим, столько кластеров получим.

Алгоритм Краскала (Kruskal)

Kruskal$(G = \langle V, E \rangle, w : E \to \mathbb{R})$:

1. Отсортировать рёбра G по возрастанию веса, инициализировать подграф $S \subseteq G$, $S := \emptyset$.
2. Для каждого ребра e в порядке возрастания веса:
   а) если конечные точки e ещё не связаны в S, добавить e в S.

1. на каждом шаге выбираем ребро с минимальным весом,
2. если оно соединяет два дерева, добавляем его в остовное дерево, если нет, пропускаем.

Алгоритм Борувски (Boruvka)

Boruvka($G = \langle V, E \rangle, w : E \to \mathbb{R}$):

1. Инициализируем список из n деревьев L, в каждом дереве по одной вершине.
2. Пока в L больше одного дерева:
   а) для каждого $T \in L$ найти ребро минимального веса, соединяющее T с $G \setminus T$;
   б) добавить все эти рёбра к минимальному остовному дереву;
   в) объединить пары пересекающихся деревьев в L (размер L при этом уменьшается вдвое).

1. можно строить минимальное остовное дерево, начав с одной вершины и добавляя ребра минимального веса, пока не покроем весь граф (алгоритм Прима, Prim's algorithm);
2. будем делать то же самое, но во всех вершинах одновременно (распараллелим этот процесс).

Алгоритм k-средних (k-means)

1. инициализировать центры кластеров каким-нибудь начальным разбиением;
2. классифицировать точки по ближайшему к ним центру кластера;
3. перевычислить каждый из центров;
4. если ничего не изменилось, остановиться, если изменилось повторить.

kMeans(X,|C|):

1. Инициализировать центры |C| кластеров:

$$\mu_1, \ldots, \mu_{|C|}.$$

2. Пока принадлежность кластерам не перестанет изменяться:
   а) определить принадлежность x_i к кластерам:

$$\text{clust}_i := \text{argmin}_{c \in C} \rho(x_i, \mu_c);$$

   б) определить новое положение центров:

$$\mu_c := \frac{\sum_{\text{clust}_i = c} f_j(x_i)}{\sum_{\text{clust}_i = c} 1}.$$

Алгоритм k-средних (k-means)

Сначала определяется **центр кластера**, а затем группируют все объекты в пределах заданного от центра порогового значения.

Недостатки:

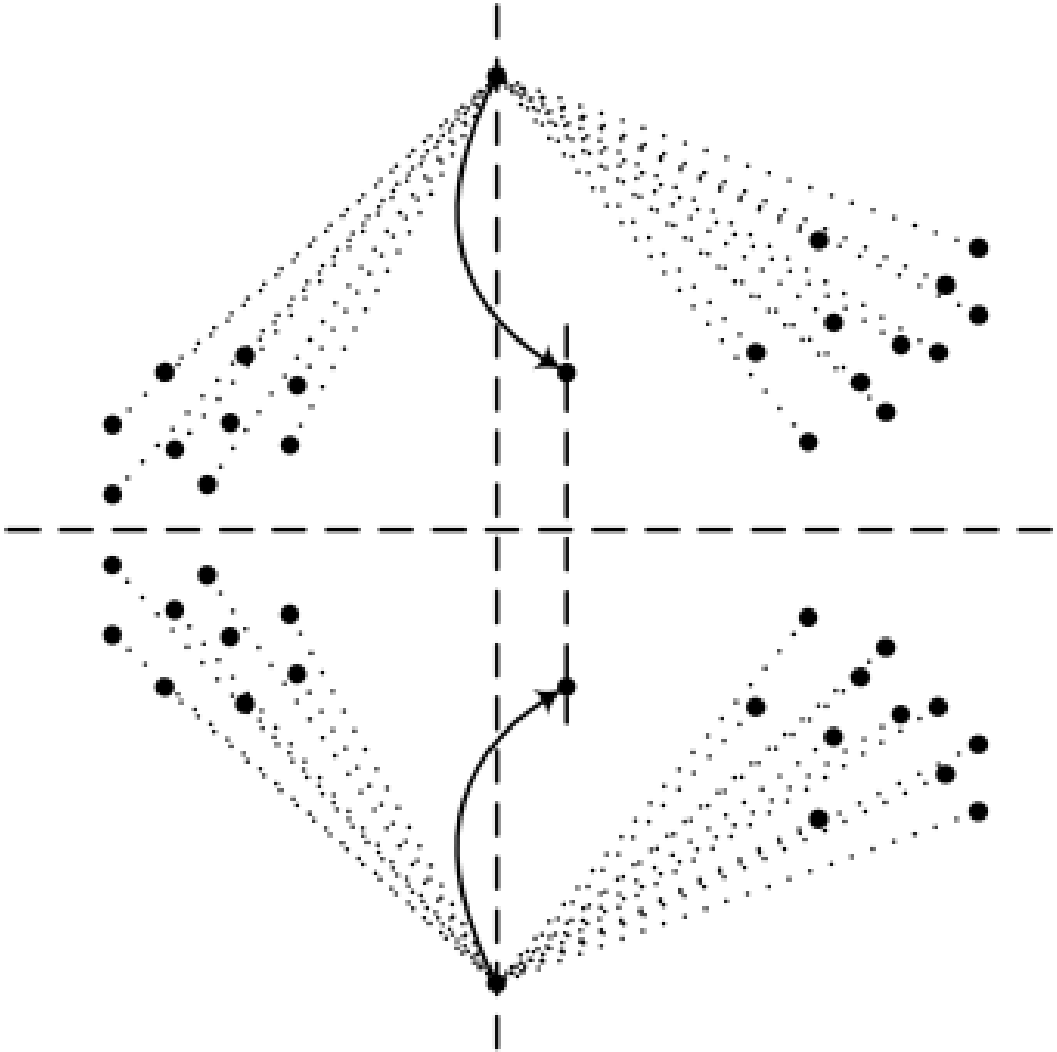
- Чувствительность к выбросам
- Необходимо заранее задавать количество кластеров, а не как в иерархическом анализе, получать это в качестве результата

Проблему с выбором числа кластеров можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров.

Достоинства:

- Простота использования
- В качестве метрики используется Евклидово расстояние
- Возможность наглядной интерпретации кластеров с использованием графика «Средних значений в кластерах»

Пример некорретной кластеризации



Алгоритмы нечеткой кластеризации

Мера принадлежности кластеру - вещественное число из [0,1], и точки на краю кластера sменьше принадлежатї кластеру, чем в центре. Будем обозначать принадлежность кластеру $c \in C$ через $u_c(x)$. Меры принадлежности обычно выбирают так, чтобы

$$\forall x \; u_c(x) \geqslant 0 \qquad \sum_{c \in C} u_c(x) = 1$$

Нечеткие алгоритмы кластеризации (одним из которых является алгоритм с-средних) минимизируют ту или иную меру ошибки. Часто применяется мера

$$E(C) = \sum_{c \in C} \sum_{x \in X} u_c^m(x) \rho^2 \left(x, \mathrm{Center}_c\right)$$

где m - некоторый вещественный параметр.

Алгоритмы нечеткой кластеризации

cMeans(X,|C|):

1. Случайно выбрать коэффициенты $u_c(x)$ для всех $x \in X$ и $c \in C$.
2. Пока алгоритм не сойдётся:
   а) Для всех $c \in C$

$$\text{Center}_c := \frac{\sum_{x \in X} u_c(x)^m x}{\sum_{x \in X} u_c(x)^m}.$$

   б) Для всех $c \in C$ и всех $x \in X$

$$u_c(x) := \frac{1}{\sum_{c' \in C} \left(\frac{\rho(\text{Center}_c, x)}{\rho(\text{Center}_{c'}, x)} \right)^{2/(m-1)}}.$$

- при $m = 2$, то перевзвешивание эквивалентно линейной нормализации коэффициентов так, чтобы их сумма была равна единице.
- при $m \to 1$ все больший и больший вес придается самому близкому кластеру, и алгоритм становится все более похож на алгоритм k-средних.

Этапы кластерного анализа

Принятие решений о числе кластеров

1. Необходимо руководствоваться практическими и теоретическими соображениями. Исходя из цели исследования, например, может быть необходимо три кластера.

2. В иерархической кластеризации в качестве критерия используются расстояния. Необходимо смотреть на **коэффициент в протоколе объединения** (расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учётом предусмотренного преобразования значений).

   - Когда мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить. Иначе будут объединены кластеры, находящиеся на большом расстоянии друг от друга.

   - Оптимальным считается число кластеров равное разности количества наблюдений и количества шагов, после которого коэффициент увеличивается скачкообразно.

3. Размеры кластеров должны быть значимыми.

Оценка качества кластеризации

- Необходимо выполнять кластерный анализ одних и тех же данных, но с использованием **различных способов измерения расстояния**.
- Сравнить результаты, полученные на основе различных способов расстояния, чтобы определить, насколько совпадают полученные результаты.
- Разбить данные на **две равные части** случайным образом. Выполнить кластерный анализ отдельно для каждой половины. Сравнить кластерные центроиды двух подвыборок.
- Случайным образом **удалить некоторые переменные**. Выполнить кластерный анализ по сокращенному набору переменных. Сравнить результаты с полученными на основе полного набора переменных.

Пример 1. Тестирование алгоритма k-means

Исходный набор данных

№	Идентификатор	Координата x	Координата y	Координата z	№	Идентификатор	Координата x	Координата y	Координата z
1	p101	0,20	0,20	0,20	26	p203	0,61	0,65	0,49
2	p201	0,30	0,40	0,40	27	p304	0,60	0,66	0,67
3	p301	0,60	0,60	0,60	28	p305	0,55	0,60	0,72
4	p401	0,80	0,80	0,80	29	p306	0,68	0,70	0,60
5	p102	0,14	0,23	0,22	30	p307	0,59	0,54	0,69
6	p103	0,20	0,25	0,20	31	p308	0,71	0,62	0,55
7	p104	0,22	0,22	0,24	32	p309	0,50	0,59	0,51
8	p105	0,25	0,14	0,21	33	p310	0,56	0,68	0,59
9	p106	0,29	0,21	0,26	34	p311	0,67	0,71	0,61
10	p107	0,26	0,23	0,17	35	p402	0,82	0,84	0,97
11	p108	0,17	0,26	0,18	36	p403	0,81	0,85	0,77
12	p109	0,28	0,12	0,26	37	p404	0,80	0,86	0,73
13	p110	0,15	0,13	0,19	38	p405	0,87	0,80	0,79
14	p202	0,31	0,46	0,46	39	p406	0,88	0,78	0,83
15	p203	0,30	0,56	0,49	40	p407	0,92	0,86	0,80
16	p204	0,35	0,48	0,43	41	p408	0,71	0,89	0,92
17	p205	0,34	0,40	0,42	42	p409	0,83	0,96	0,87
18	p206	0,33	0,43	0,41	43	p410	0,84	0,90	0,89
19	p207	0,39	0,47	0,44	44	p411	0,98	0,91	0,91
20	p208	0,38	0,49	0,50	45	p412	0,93	0,87	0,85
21	p209	0,37	0,42	0,42	46	p413	0,91	0,85	0,82
22	p210	0,32	0,44	0,51	47	p414	0,79	0,81	0,96
23	p211	0,25	0,45	0,52	48	p415	0,77	0,76	0,83
24	p212	0,20	0,39	0,38	49	p416	0,99	0,75	0,96
25	p302	0,62	0,64	0,57					

Пример 1. Тестирование алгоритма k-means

Средние значения координат всей статистической совокупности

	v1	v2	v3
Средние значения	0,538	0,569	0,568

Ковариационная матрица статистической совокупности

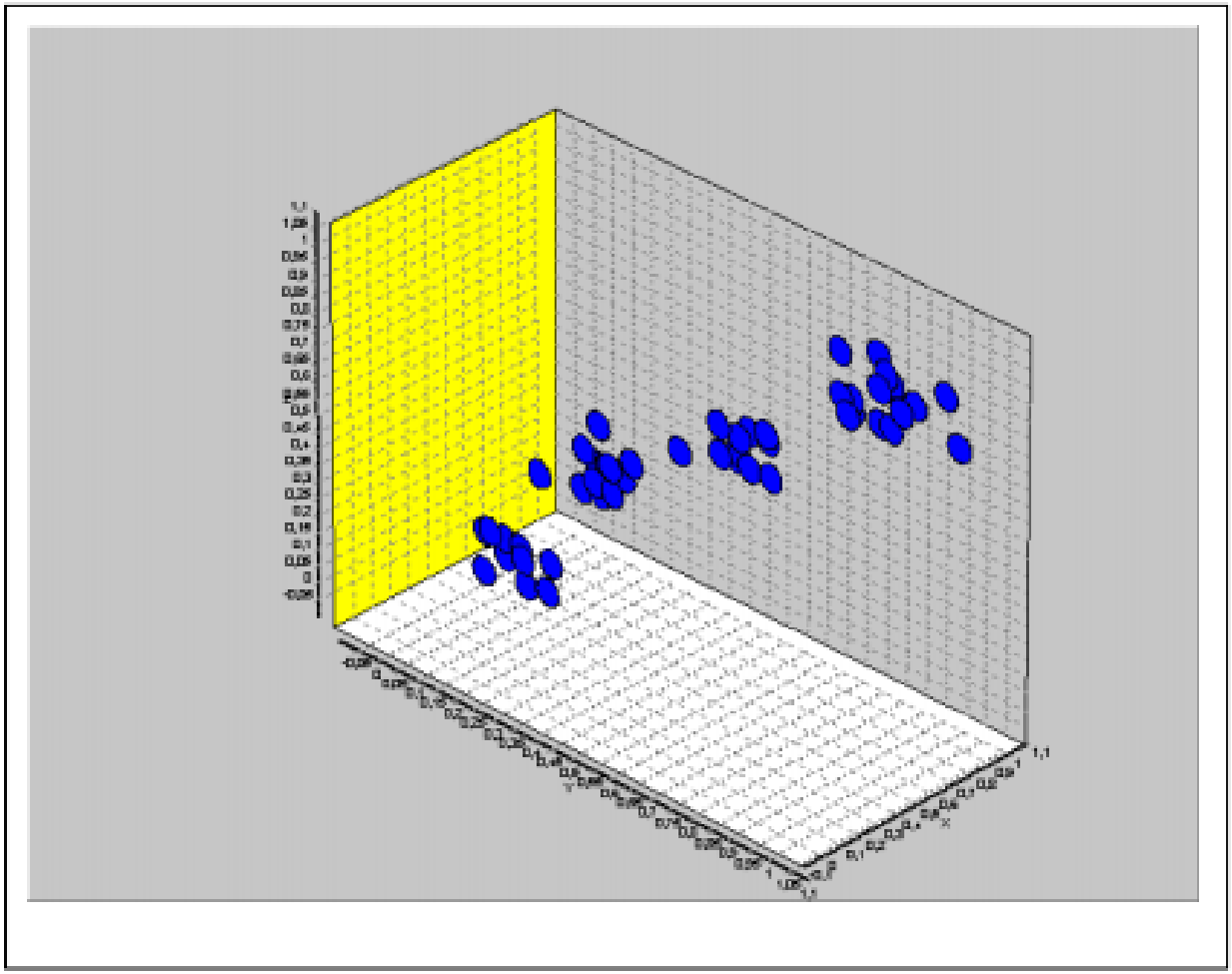
	v1	v2	v3
v1	0,070	0,060	0,060
v2	0,060	0,060	0,057
v3	0,060	0,057	0,060

Среднеквадратические отклонения

	v1	v2	v3
Среднеквадратические отклонения	0,265	0,244	0,245

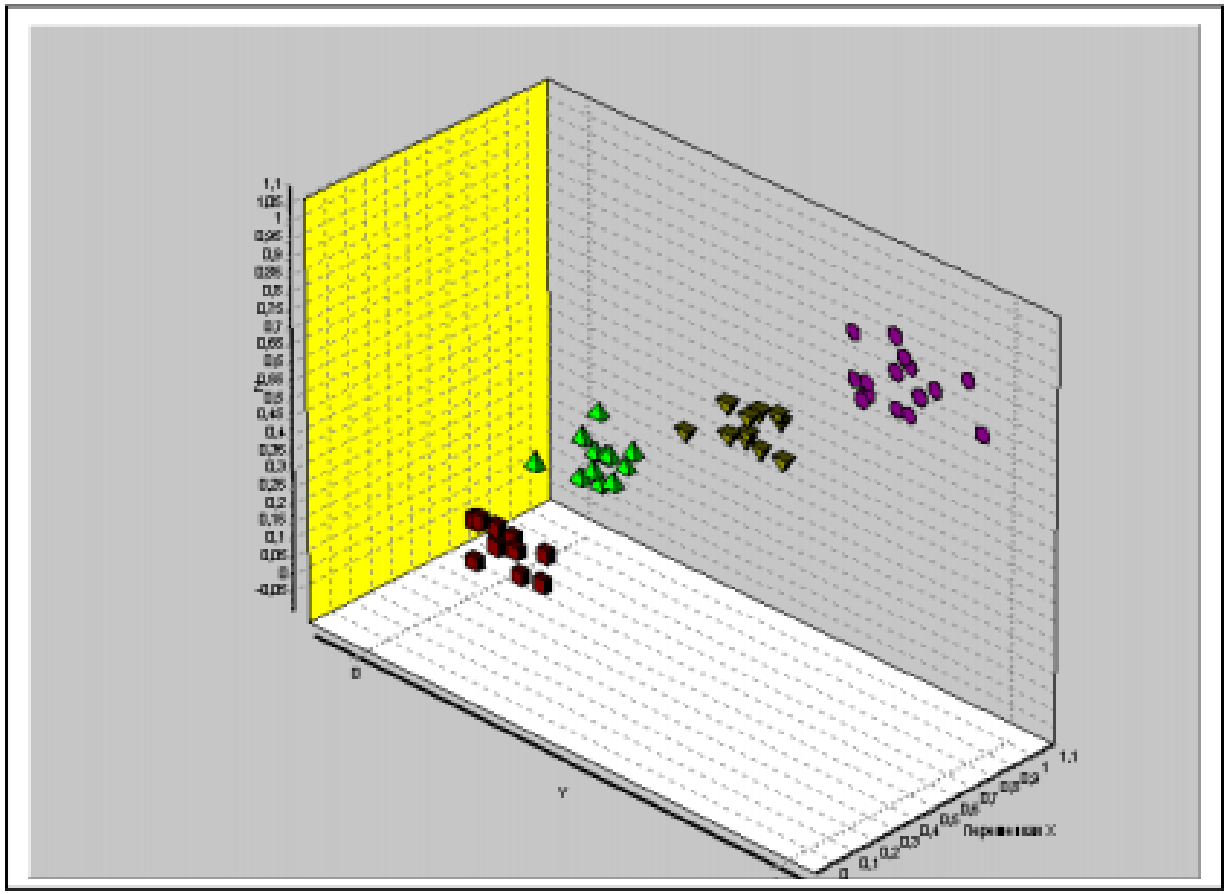
Пример 1. Тестирование алгоритма k-means

Взаимное расположение данных в пространстве



Пример 1. Тестирование алгоритма k-means

Результаты работы алгоритма k-means



Пример 1. Тестирование алгоритма k-means

Расстояния между кластерами

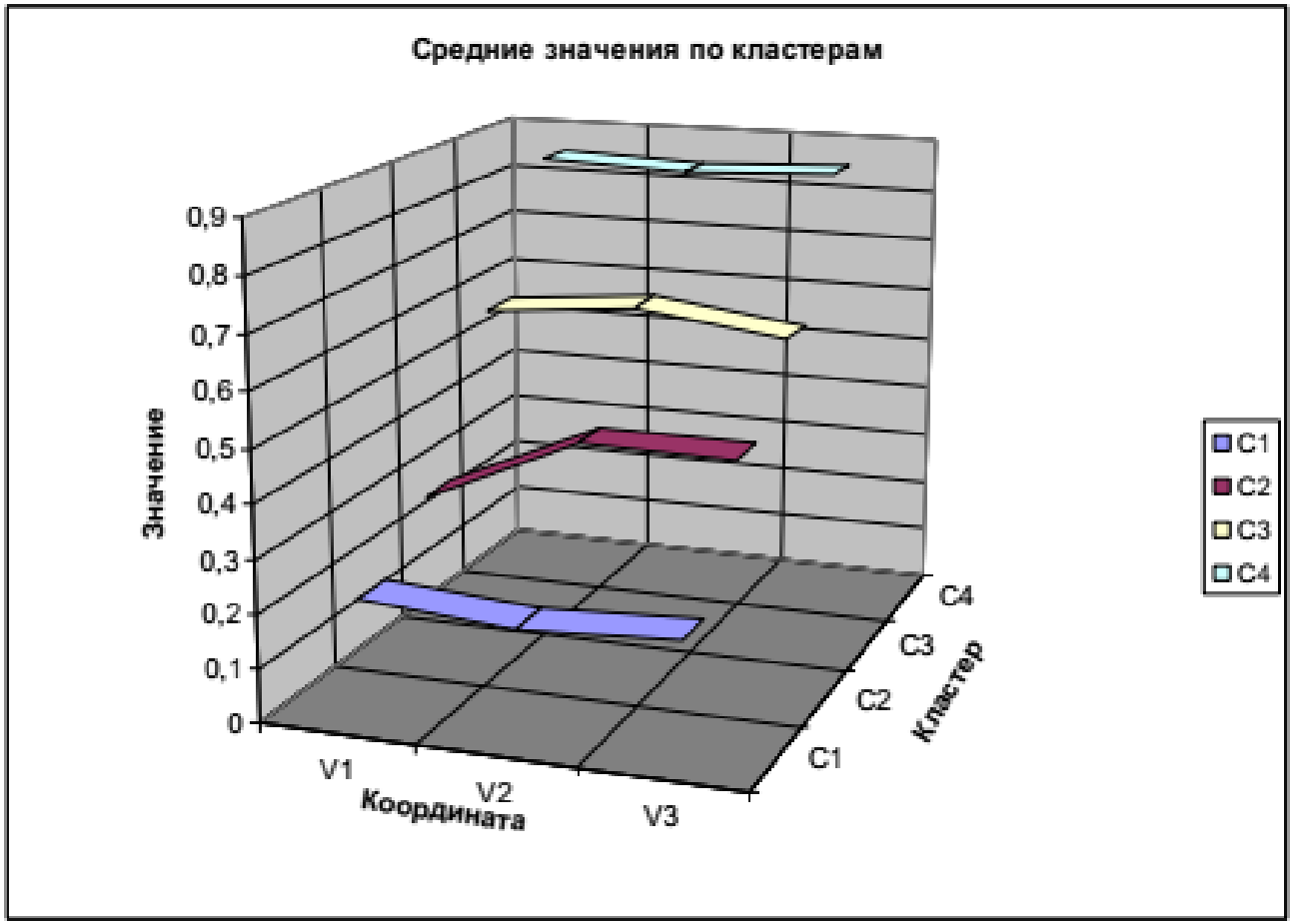
Кластеры	№ 1	№ 2	№ 3	№ 4
№ 1		0,1378	0,5228	1,304
№ 2	0,3712		0,1468	0,6352
№ 3	0,7231	0,3832		0,1778
№ 4	1,142	0,797	0,4217	

Средние значения координат в кластерах

				№ 1	№ 2	№ 3	№ 4
1	2	3	4	5	6	7	8
		Кол-во точек		10	12	11	16
		V1		0,216	0,32	0,6082	0,8531
		V2		0,199	0,4492	0,6355	0,8431
		V3		0,213	0,4483	0,6	0,8563
		Имя кластера					

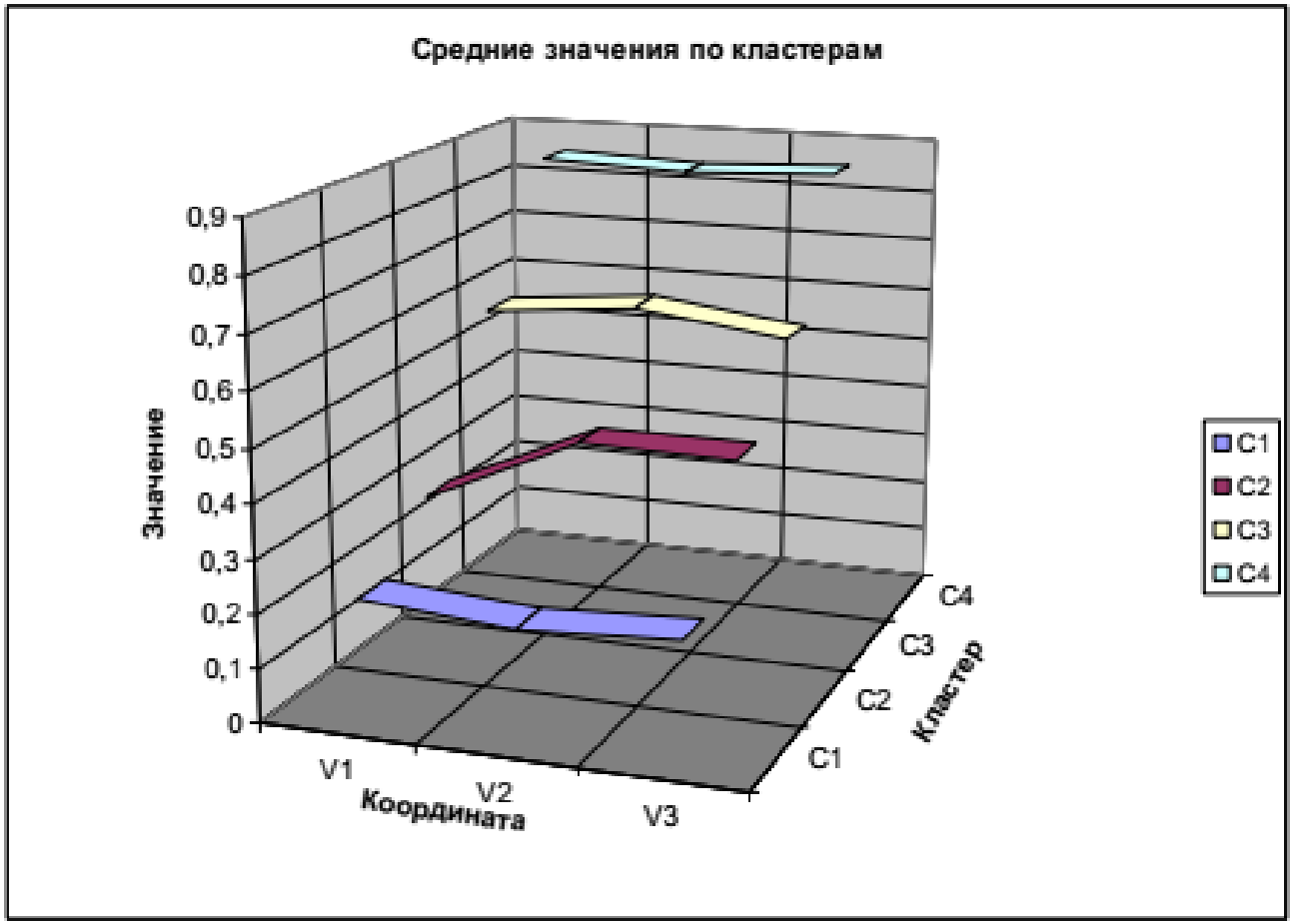
Пример 1. Тестирование алгоритма k-means

Средние значения координат в кластерах



Пример 1. Тестирование алгоритма k-means

Средние значения координат в кластерах



Пример 1. Тестирование алгоритма k-means

Критерий Фишера проверки гипотезы о различии средних значений к кластерах

Пере-мен-ная	Межклас-терная SS	Число степеней свободы	Внутриклас-терная SS	Число степеней свободы	Значение Критерия Фишера	Уровень значимости
1	2	3	4	5	6	7
V1	3,316	3	0,1893	44	256,9	0,036%
V2	3,03	3	0,1348	44	329,7	0,025%
V3	2,947	3	0,174	44	248,3	0,0379%

Пример 2. Исследование структуры личных подсобных хозяйств

- В выборку вошли **29010** хозяйств Саратовской области.
- Кластерный анализ проводился с помощью алгоритма k-средних.
- Разбиение проводилось на 16, 24 и 32 кластера.
- Из множества показателей, содержащихся в переписном листе 3 ЛПХ, выделены 43 наиболее значимых (по мнению экспертов) показателя.

Условные названия групп показателей:

- A – трудовые ресурсы (показатели V1, V2);
- B – площади земли (показатели V3..V6);
- C – площади посевов (показатели V7..V12);
- D – деревья, кусты и ягодники (показатели V13..V18);
- E – скот (показатели V19..V26, V32);
- F – птица (показатели V27..V31);
- G – места для содержания скота и птицы (показатели V33..V37);
- H – техника (показатели V38..V43).

Обозначение показателей

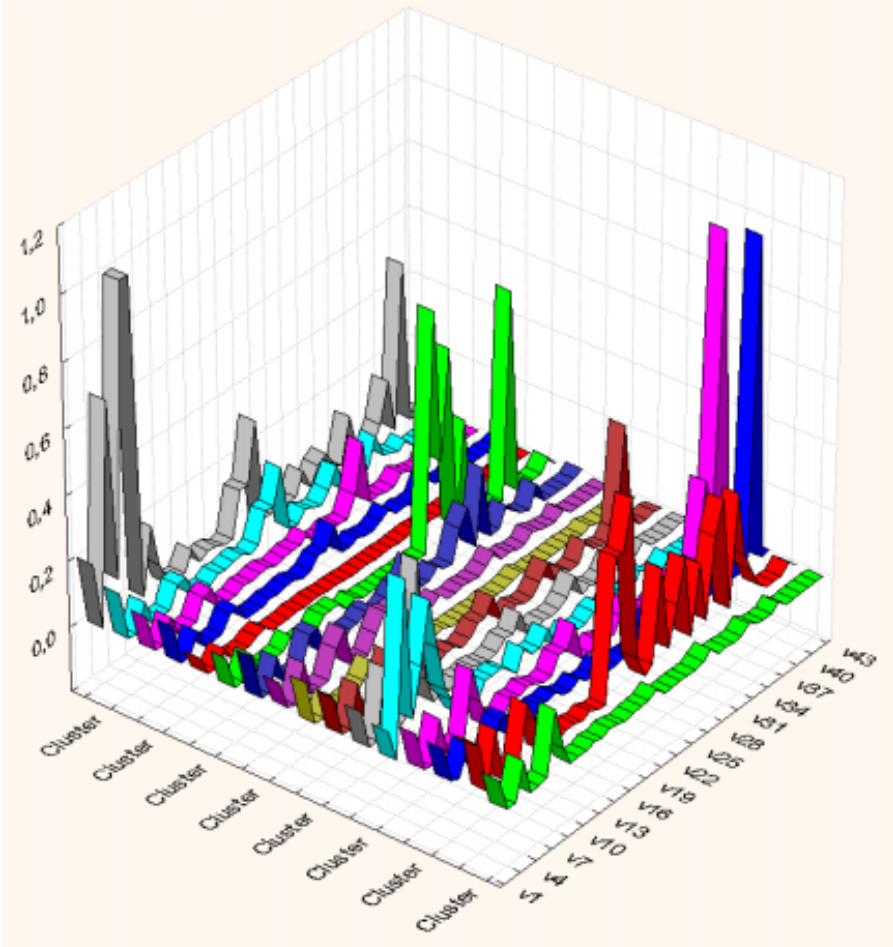
A	1	v1	Число членов семьи, занятых в хозяйстве, чел	ЧЧлСем
	2	v2	Численность сезонных работников,чел	ЧислСезРаб
B	3	v3	Общая прощадь земли в собственности, га	ПлЗемСобств
	4	v4	Приусадебный земельный участок, га	ПрЗемУч
	5	v5	Полевые земельные участки, га	ПоЗемУч
	6	v6	Общая прощадь используемой земли,чел	ОбщИспЗем
C	1	v7	Всего посевов по урожай, кв.м.	ОбщПосУр
	2	v8	в т.ч. картофель, кв.м.	КарПосУр
	3	v9	Овощные и бахчевые культуры отр. гр., кв.м.	ОвОткрГр
	4	v10	Овощи закрытого грунта, кв.м.	ОвЗакрГр
	5	v11	Кормовые культуры, кв.м.	КормКульт
	6	v12	Площадь посевов за пределом уч-ка, кв.м.	ПлЗаПредУч
D	1	v13	Яблони, шт	Ябл
	2	v14	Груши, шт	Груша
	3	v15	Слива, шт	Слива
	4	v16	Вишня, шт	Виншя
	5	v17	Земляника и клубника, кв.м.	ЗемлКлубн
	6	v18	Смородина, шт	Смор
E	1	v19	Крупный рогатый скот,голов	КРС
	2	v20	КРС молочного стада, голов	КРСмол
	3	v21	из него: коровы, голов	КРСмолКор
	4	v22	КРС мясного стада, голов	КРМмяс
	5	v23	из него: коровы, голов	КРСмясКор
	6	v24	Свиньи, голов	Свиньи
	7	v25	Овцы, голов	Овцы
	8	v26	Козы, голов	Козы
F	1	v27	Птица, голов	Птица
	2	v28	Куры яичного направления, голов	КурЯич
	3	v29	Куры мясного направления, голов	КурМяс
	4	v30	Утки, голов	Утки
	5	v31	Гуси, голов	Гуси
	6	v32	Лошади, голов	Лош
G	1	v33	Мест для содержания КРС, шт	МСКРС
	2	v34	Мест для содержания свиней, шт	МССвин
	3	v35	Мест для содержания овец и коз, шт	МСОвКоз
	4	v36	Мест для содержания лошадей, шт	МСЛош
	5	v37	Мест для содержания птицы, шт	МСПтиц
H	1	v38	Тракторы, шт	Тракт
	2	v39	Плуги тракторные, шт	ПлТракт
	3	v40	Косилки тракторные, шт	КосТракт
	4	v41	Мотоблоки, мотокультиваторы, шт	Мблок
	5	v42	Автомобили грузопассажирские, шт	АвтГрПасс
	6	v43	Установки доильные, шт	УстДоил

Кластеры и средние значения показателей

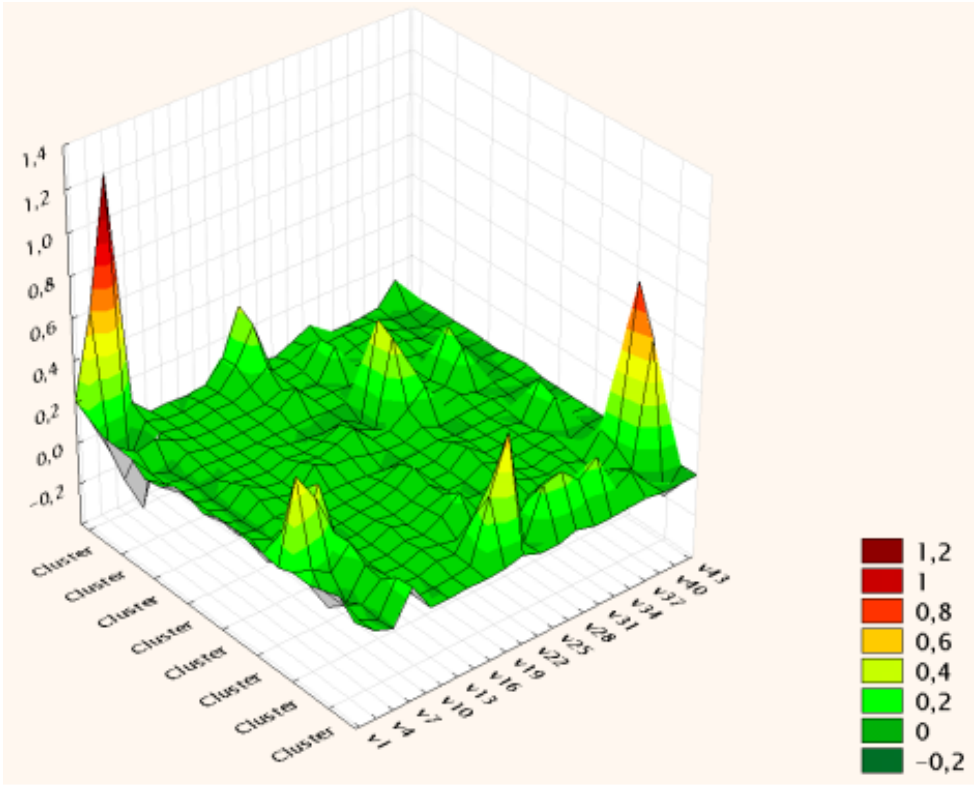
				c1	c2	c3	c4	c5	c6	c7	c8
				2767	14	82	38	4	723	141	6658
A	1	v1	ЧЧлСем	2,04	3,43	2,43	2,55	2,00	2,41	2,50	2,69
	2	v2	ЧислСезРаб	0,02	0,79	0,06	0,05	0,00	0,01	0,01	0,01
B	1	v3	ПлЗемСобств	0,15	0,22	0,12	0,17	47,12	0,28	0,16	0,07
	2	v4	ПрЗемУч	0,15	0,22	0,12	0,17	0,12	0,27	0,16	0,07
	3	v5	ПоЗемУч	0,00	0,00	0,00	0,00	14,00	0,01	0,00	0,00
	4	v6	ОбщИспЗем	0,15	0,22	0,12	0,17	24,86	0,28	0,16	0,07
C	1	v7	ОбщПосУр	1072,28	1179,14	627,60	1197,55	25416,50	2386,78	875,96	270,98
	2	v8	КарПосУр	940,67	850,14	446,89	974,97	512,50	2201,13	715,02	128,65
	3	v9	ОвОткрГр	126,50	221,86	179,33	186,50	118,00	172,65	158,43	141,60
	4	v10	ОвЗакрГр	0,01	0,00	0,48	0,00	0,00	0,01	0,00	0,20
	5	v11	КормКульт	5,09	107,14	0,90	36,08	36,25	13,00	2,51	0,52
	6	v12	ПлЗаПредУч	1,16	0,00	0,00	0,16	0,00	1,38	0,00	0,16
D	1	v13	Ябл	1,11	1,00	3,99	2,32	4,00	1,34	1,70	1,12
	2	v14	Груша	0,07	0,14	0,45	0,26	0,50	0,16	0,16	0,10
	3	v15	Слива	0,50	0,07	1,45	1,08	1,50	0,64	0,79	0,53
	4	v16	Вишня	0,87	0,86	1,88	1,42	2,00	0,94	1,72	0,95
	5	v17	ЗемлКлубн	2,13	1,29	13,00	5,45	1,50	1,95	4,26	3,39
	6	v18	Смор	1,98	0,00	8,28	3,13	12,50	2,43	4,02	2,75
E	1	v19	КРС	0,69	22,07	0,96	3,05	1,50	1,88	2,53	0,04
	2	v20	КРСмол	0,69	21,64	0,96	3,05	1,50	1,88	2,49	0,04
	3	v21	КРСмолКор	0,36	6,29	0,48	1,32	0,75	0,90	1,08	0,00
	4	v22	КРМмяс	0,00	0,43	0,00	0,00	0,00	0,00	0,04	0,00
	5	v23	КРСмясКор	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	6	v24	Свиньи	0,60	2,14	0,73	2,61	3,00	1,24	2,33	0,16
	7	v25	Овцы	0,17	46,00	0,06	1,45	0,00	0,1148	1,11	0,04
	8	v26	Козы	0,07	0,71	0,06	0,00	0,00	0,16	0,14	0,04
F	1	v27	Птица	6,38	76,50	19,28	21,71	10,00	14,38	18,67	3,91
	2	v28	КурЯич	4,13	35,71	9,29	9,82	3,75	9,86	10,82	2,54
	3	v29	КурМяс	1,16	7,50	5,28	7,26	1,00	1,79	2,93	0,79
	4	v30	Утки	0,68	22,36	3,21	3,29	3,75	1,59	2,70	0,41
	5	v31	Гуси	0,36	10,21	1,50	0,29	1,50	0,89	2,16	0,15
	6	v32	Лош	0,04	2,71	0,01	0,13	0,00	0,12	0,11	0,00
G	1	v33	МСКРС	1,34	18,79	1,26	4,00	2,25	2,63	3,33	0,20
	2	v34	МССвин	1,59	2,86	1,67	5,03	4,75	2,31	3,81	0,60
	3	v35	МСОвКоз	0,48	46,14	0,66	1,39	0,00	1,07	1,93	0,18
	4	v36	МСЛош	0,06	3,00	0,01	0,13	0,00	0,16	0,13	0,01
	5	v37	МСПтиц	15,77	59,86	21,98	30,24	18,75	24,40	26,33	9,17
H	1	v38	Тракт	0,00	0,07	0,01	0,82	0,00	0,01	1,01	0,00
	2	v39	ПлТракт	0,00	0,00	0,01	0,50	0,00	0,00	0,14	0,00
	3	v40	КосТракт	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00
	4	v41	Мблок	0,00	0,00	1,00	0,11	0,00	0,00	0,00	0,00
	5	v42	АвтГрПасс	0,01	0,21	0,05	0,11	0,00	0,04	0,07	0,02
	6	v43	УстДоил	0,00	0,00	0,01	0,03	0,00	0,00	0,00	0,00
			Имя кластера	CBF	EGF	HCD	HDC	DB	CBE	HEC	OOC

				c9	c10	c11	c12	c13	c14	c15	c16
				1359	634	3	11373	2599	1422	1191	2
A	1	v1	ЧЧлСем	1,98	2,78	2,00	1,29	2,47	2,30	3,00	4,00
	2	v2	ЧислСезРаб	0,06	0,01	0,00	0,03	0,01	0,02	0,01	0,00
B	1	v3	ПлЗемСобств	0,15	0,12	0,09	0,06	0,08	0,09	0,11	60,08
	2	v4	ПрЗемУч	0,15	0,12	0,09	0,06	0,08	0,09	0,11	0,15
	3	v5	ПоЗемУч	0,00	0,00	0,00	0,00	0,00	0,00	0,00	59,93
	4	v6	ОбщИспЗем	0,15	0,12	0,09	0,06	0,08	0,09	0,11	60,08
C	1	v7	ОбщПосУр	678,14	644,21	104,33	130,91	342,57	451,24	466,58	1150,00
	2	v8	КарПосУр	408,79	442,38	30,00	82,79	230,94	301,46	318,46	1000,00
	3	v9	ОвОткрГр	258,80	197,50	74,33	47,61	109,19	148,37	143,97	150,00
	4	v10	ОвЗакрГр	0,33	0,12	0,00	0,08	0,09	0,13	0,09	0,00
	5	v11	КормКульт	10,21	4,21	0,00	0,41	2,36	1,28	4,06	0,00
	6	v12	ПлЗаПредУч	0,29	0,00	0,00	0,12	0,62	1,14	0,00	0,00
D	1	v13	Ябл	3,80	2,11	2,67	0,54	0,86	1,35	1,43	4,50
	2	v14	Груша	1,06	0,28	0,33	0,03	0,07	0,14	0,13	0,00
	3	v15	Слива	2,56	1,29	1,00	0,19	0,48	0,59	0,63	0,50
	4	v16	Вишня	5,30	2,62	0,00	0,43	0,95	1,20	1,47	1,00
	5	v17	ЗемлКлубн	19,17	6,90	10,67	0,89	2,52	3,60	4,04	0,00
	6	v18	Смор	25,12	5,85	0,00	0,95	2,46	2,91	4,01	0,50
E	1	v19	КРС	0,64	2,88	0,00	0,02	2,05	1,06	5,46	7,00
	2	v20	КРСмол	0,63	2,84	0,00	0,02	2,05	1,05	5,43	7,00
	3	v21	КРСмолКор	0,30	1,15	0,00	0,01	1,12	0,45	2,30	3,50
	4	v22	КРМмяс	0,00	0,02	0,00	0,00	0,01	0,01	0,03	0,00
	5	v23	КРСмясКор	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00
	6	v24	Свиньи	0,72	3,27	0,00	0,06	1,34	1,36	4,05	1,00
	7	v25	Овцы	0,38	1,41	0,00	0,02	0,55	0,51	2,82	0,00
	8	v26	Козы	0,04	0,10	0,00	0,02	0,02	0,06	0,21	0,00
F	1	v27	Птица	8,91	49,18	369,00	1,06	7,75	31,29	20,88	33,50
	2	v28	КурЯич	5,82	18,95	33,33	0,71	4,84	27,13	12,98	0,00
	3	v29	КурМяс	1,69	4,32	233,33	0,21	1,37	1,88	3,00	20,00
	4	v30	Утки	0,90	23,45	2,33	0,10	1,02	1,09	2,73	7,00
	5	v31	Гуси	0,42	2,19	100,00	0,04	0,52	1,09	1,93	6,50
	6	v32	Лош	0,03	0,12	0,00	0,00	0,08	0,04	0,28	0,00
G	1	v33	МСКРС	1,26	3,53	0,00	0,19	2,68	1,68	5,66	7,00
	2	v34	МССвин	1,89	5,02	0,00	0,32	2,68	2,55	5,32	1,00
	3	v35	МСОвКоз	0,94	1,94	0,00	0,14	1,10	0,93	3,46	0,00
	4	v36	МСЛош	0,04	0,13	0,00	0,01	0,09	0,05	0,28	0,00
	5	v37	МСПтиц	16,10	49,25	440,00	4,20	14,90	39,02	24,59	33,50
H	1	v38	Тракт	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,50
	2	v39	ПлТракт	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	3	v40	КосТракт	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,50
	4	v41	Мблок	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	5	v42	АвтГрПасс	0,02	0,03	0,33	0,00	0,01	0,02	0,02	0,00
	6	v43	УстДоил	0,00	0,02	0,00	0,00	0,00	0,00	0,01	0,00
			Имя кластера	DCB	FDC	FGH	OOA	OCG	OCF	EHF	ADE

Кластеры и средние значения показателей



Кластеры и средние значения показателей



Сравнение разбиений с разным количеством показателей

№ п/п	по 37 показателям			по показателям			Совпаде-ние кластеров
	Обозначение кластеров	Количество хозяйств в кластере	Имя кластера	Обозна-чение кластеров	Кол-во хозяйств в кластере	Имя кластера	
1	C1	3	FGH	C11	3	FGH	полное
2	C2	6658	OCA	C8	6658	OOC	сущест-венное
3	C3	12534	OOC	C12	11373	OOA	почти полное
4	C4	38	HEC	C4	38	HDC	полное*
5	C5	3245	OCB	C 1	2767	CBF	сущест-венное
6	C6	1393	EFG	C15	1191	EHF	сущест-венное
7	C7	3030	OCE	C13	2599	OCG	хорошее
8	C8	757	CBE	C6	723	CBE	почти полное
9	C9	1	HCB	-	-	-	уникаль-ное хозяй-ство
10	C10	35	HEF	-	-	-	уникаль-ное хозяй-ство
11	C11	6	CBE	C5	4	CDB	хорошее
12	C12	32	EFG	C2	14	EGF	хорошее по кач. пок.
13	C13	2	BAE	C16	2	ADE	полное
14	C14	10	EGH	-	-	-	уникаль-ное
15	C15	1285	OFC	C15	1191	EHF	хорошее
16	C16	81	HCA	C3	82	HCD*	полное

*) Несовпадение имен HEC и НДС обусловлено тем, что группа показателей Д исключена