Модели и задачи Data Mining

- Data Mining – совокупность большого числа различных методов обнаружения знаний.
- В современной бизнес-аналитике принято выделять в Data Mining описательные (дескриптивные) и предсказательные модели.

Кластерный анализ

1. Кластерный анализ предназначен для разбиения данных на поддающиеся интерпретации группы (1), таким образом, чтобы элементы, входящие в одну группу были максимально «схожи» (2), а элементы из разных групп были максимально «отличными» друг от друга (3).

2. Кластерный анализ – группа методов, используемых для классификации объектов или событий в относительно гомогенные (однородные) группы, которые называют кластерами (clusters).

3. Цель кластеризации - поиск существующих структур.

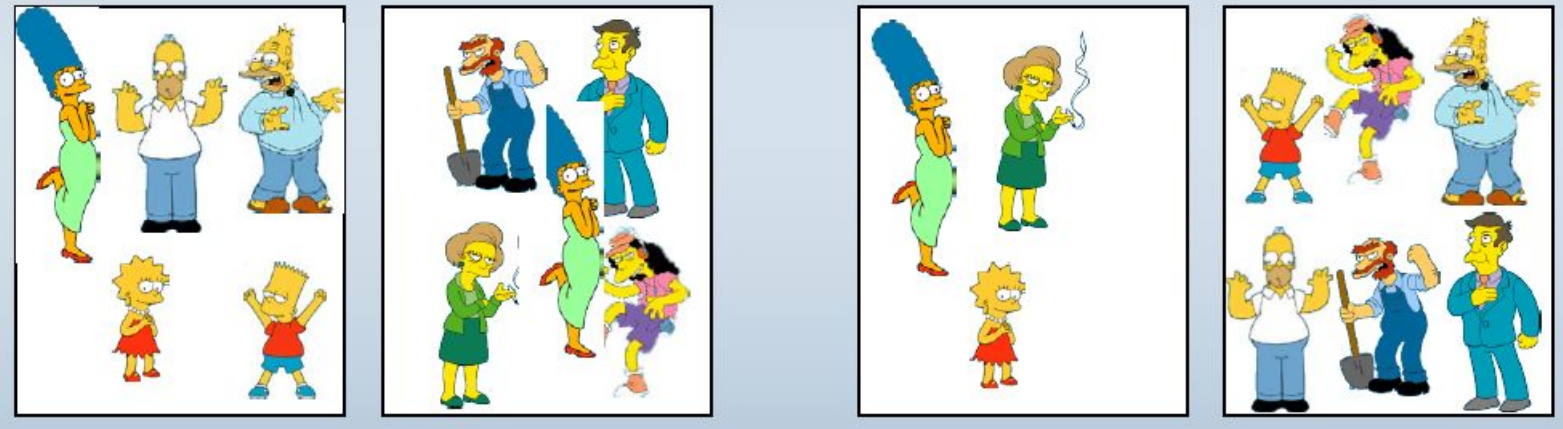
Алгоритм кластеризации — функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие идентификатор кластера $y \in Y$.

INTERNET STUDIES LAB

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
SAINT PETERSBURG

Задачи и методы кластер - анализа

Кластеризация – это разбиение элементов некоторого множества на группы на основе их схожести. Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров.



Семья | Сотрудники | Женщины | Мужчины

Лейбелинг групп – то что нужно найти

Кластеризация достаточно субъективна и зависит от цели пользователя

- В факторном анализе группируются столбцы, т.е. цель – анализ структуры множества признаков и выявление обобщенных факторов.
- В кластерном анализе–группируются строки, т.е. цель – анализ структуры множества объектов.
- Кластерный анализ выполняет классификацию объектов.
- Каждый объект (респондент) –точка в пространстве признаков.
- Задача кластерного анализа –выделение «сгущений» точек, разбиение совокупности на однородные подмножества объектов (сегментация).

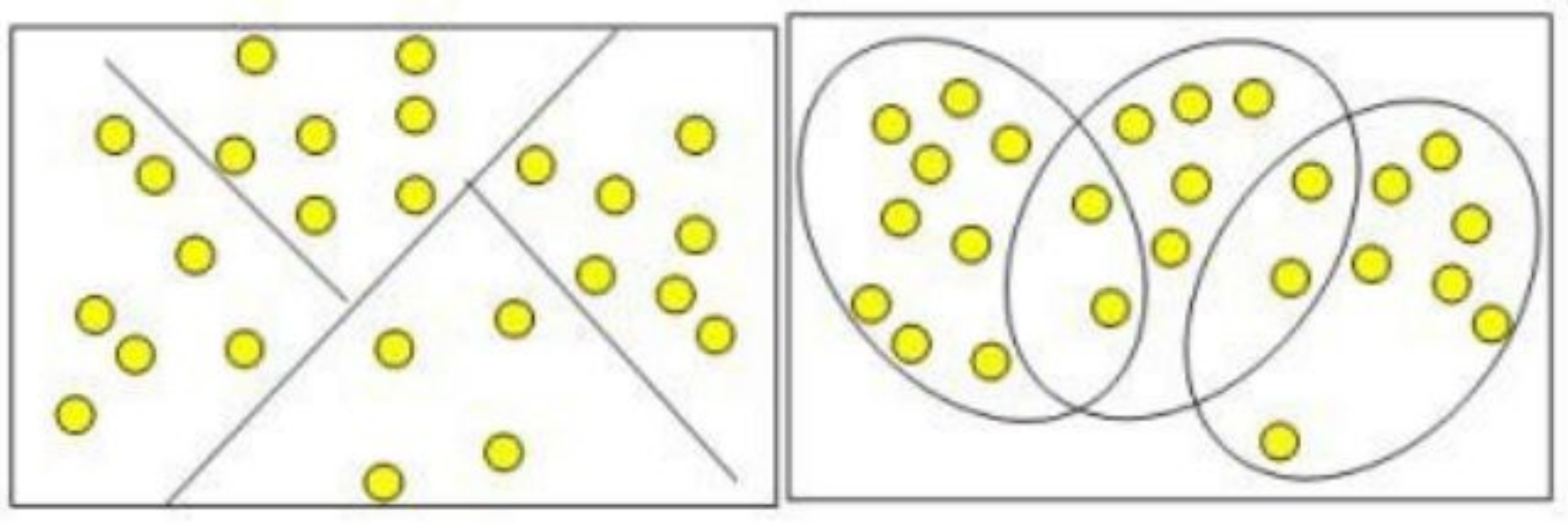
Кластерный анализ в теории

Кластерный анализ на практике

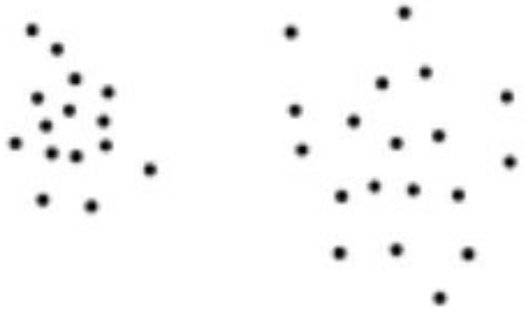
Как очертить границу кластеров?
Сколько их следует выделить?

Типы кластерных структур

- Кластеры могут быть непересекающимися (non-overlapping), и пересекающимися (overlapping).

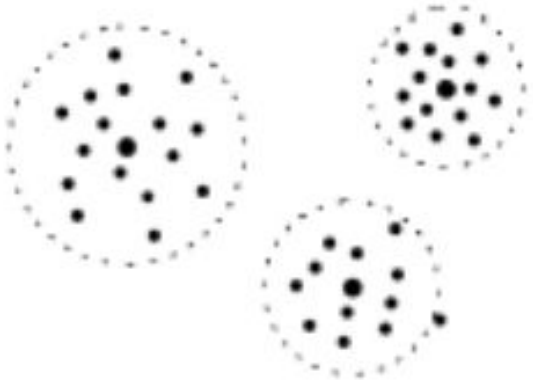


- Могут быть получены кластеры различной формы: длинными "цепочками", удлиненной формы, произвольной формы.
- Могут быть созданы кластеры определенных размеров (например, малых или крупных).



внутрикластерные расстояния, как правило, меньше межкластерных

ленточные кластеры

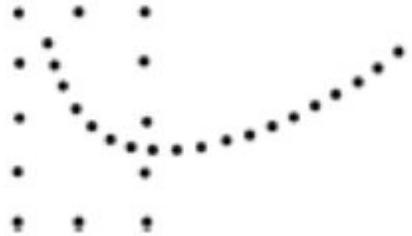


кластеры с центром

кластеры могут соединяться перемычками

кластеры могут накладываться на разреженный фон из редко расположенных объектов

кластеры могут перекрываться



кластеры могут образовываться не по сходству, а по иным типам регулярностей

кластеры могут вообще отсутствовать

- Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.
- Понятие “тип кластерной структуры” зависит от метода и также не имеет формального определения.

Выделяют четыре основных метода анализа Big Data [1]:

- Описательная аналитика (descriptive analytics) — отвечает на вопрос «Что происходит?», анализирует данные, поступающие в реальном времени, и исторические данные.
- Прогнозная или предикативная аналитика (predictive analytics) — помогает спрогнозировать наиболее вероятное развитие событий на основе имеющихся данных.
- Предписательная аналитика (prescriptive analytics) — позволяет выявить проблемные точки в любой деятельности и рассчитать, при каком сценарии их можно избежать их в будущем.
- Диагностическая аналитика (diagnostic analytics) — помогает выявлять аномалии и случайные связи между событиями и действиями.

1. Виды аналитики: описательная, прогнозная, предписывающая аналитика (projectpro.io)

Диагностическая аналитика

- Диагностическая (diagnostic) аналитика — форма расширенной аналитики, которая строится на основе описательной и анализирует данные для ответа на вопрос «Почему это произошло?», т.е. позволяет выявить факторы, оказывающие влияние на целевые параметры.

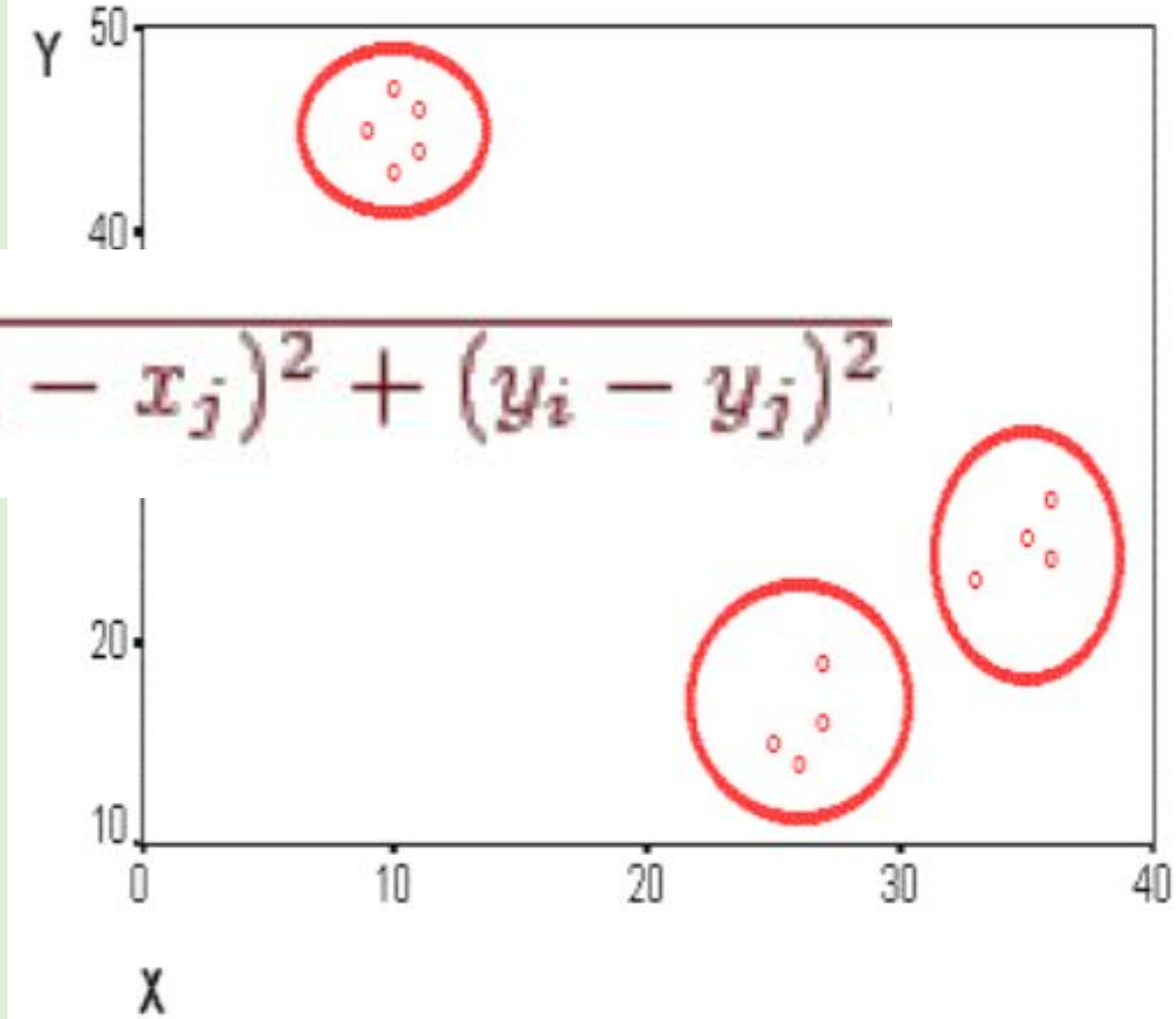
- Диагностическая аналитика позволяет изучить проблемы, определить слабые места и выявить цепочки событий.
- Методы диагностической аналитики в Loginom:
  - EM Кластеризация.
  - Кластеризация k-means, кластеризация g-means.
  - Кластеризация транзакций.

Меры сходства

- Евклидово расстояние: расстояние(x,y) = $\{\sum_i (x_i - y_i)^2\}^{1/2}$
- Квадрат евклидова расстояния чтобы придать большие веса более отдаленным друг от друга объектам расстояние (x,y) =$\sum_i (x_i - y_i)^2$
- Манхэттенское расстояние "хэмминговое" расстояние(x,y) =$\sum_i |x_i - y_i|$
- Расстояние Чебышева используют когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению: расстояние(x,y) = Максимум$|x_i - y_i|$
- Степенное расстояние для увеличения или уменьшения веса объектов при сильном отличии: расстояние(x,y) = $(\sum_i |x_i - y_i|^p)^{1/r}$
- Процент несогласия - расстояние вычисляется, если данные являются категориальными: расстояние(x,y) = (Количество $x_i \neq y_i$)/ i

Номер	Признак X	Признак Y
1	27	19
2	11	46
3	4	7
4	25	15
5	36	27
6	35	25
7	10	43
8	11	44
9	36	24
10	26	14
11	9	45
12	33	23
13	27	16
14	10	4

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$



Мера близости вычисленная как ***евклидово расстояние*** между

- Когда осей (измерений) больше, чем две, расстояние рассчитывается как: сумма квадратов разницы координат из стольких слагаемых, сколько осей (измерений) присутствует.

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

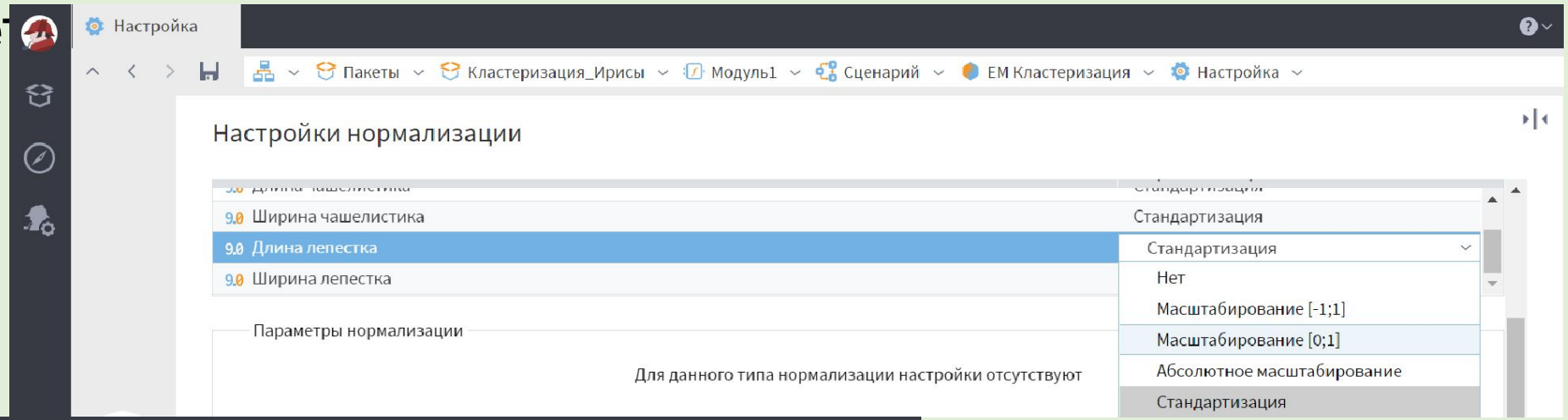
Кластер имеет следующие математические характеристики:

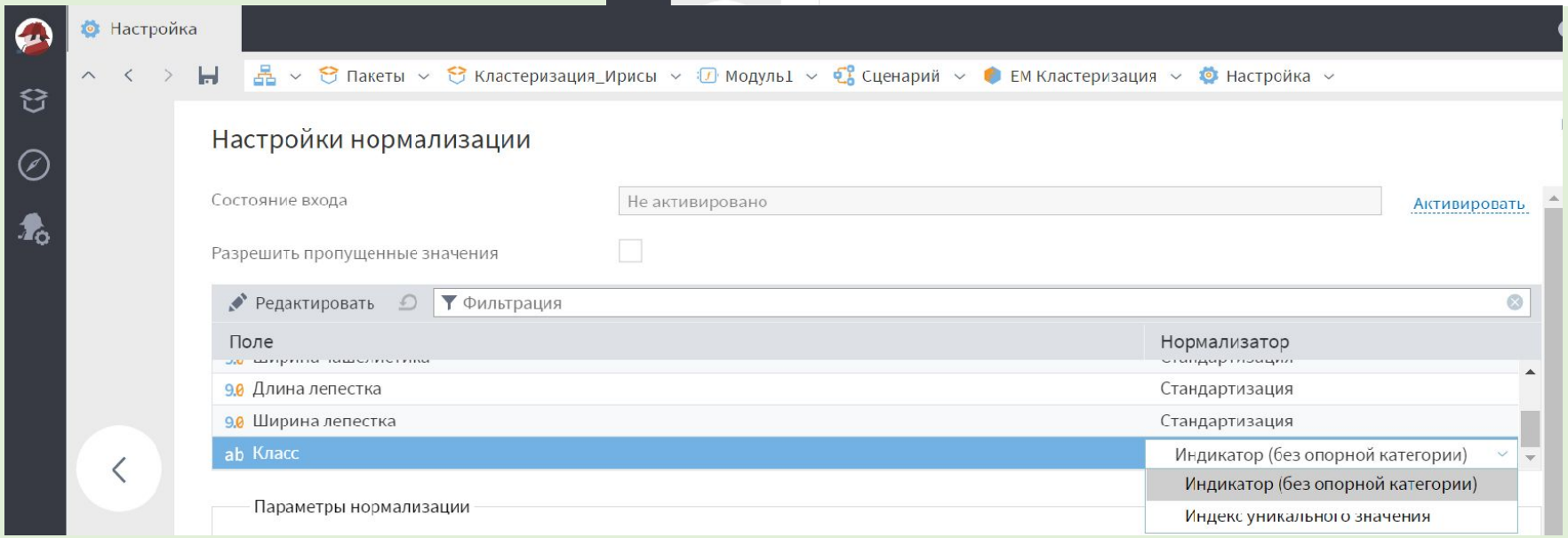
- центр,
- радиус,
- среднеквадратическое отклонение,
- размер кластера.

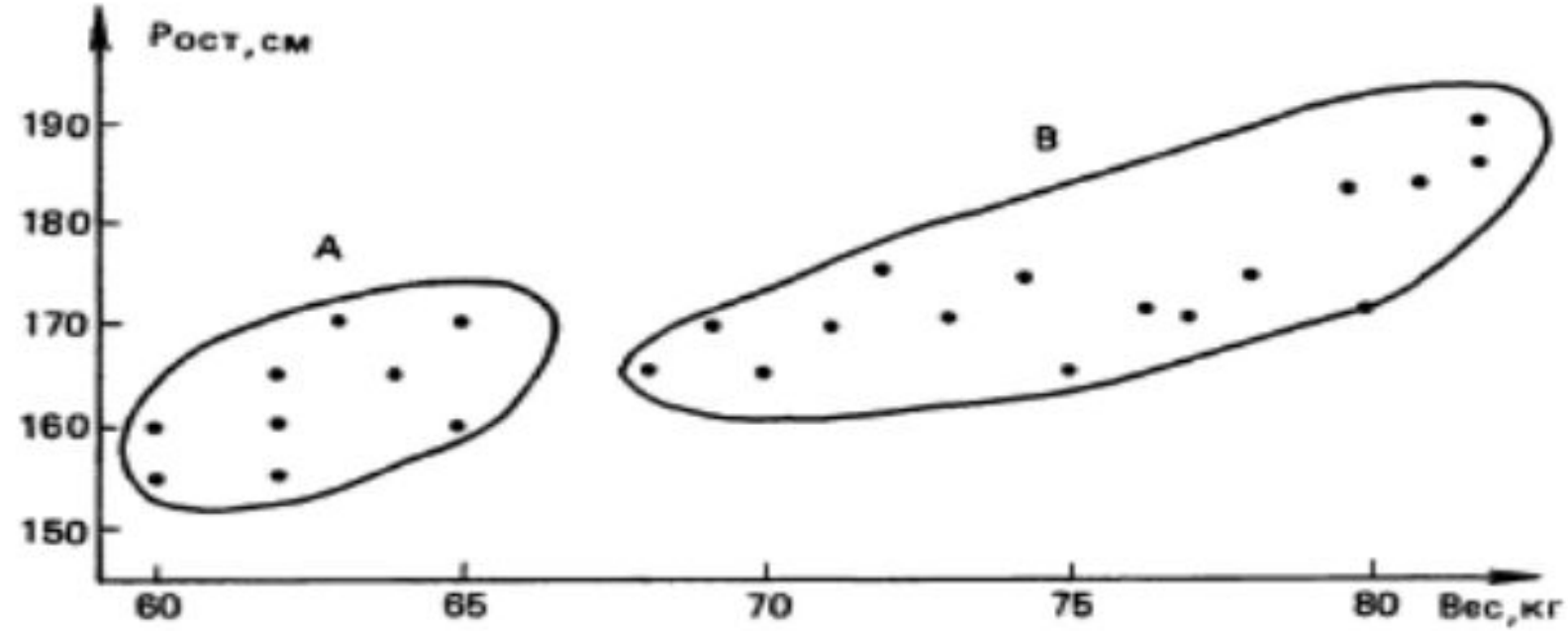
Центр кластера - это среднее геометрическое место точек в пространстве переменных.

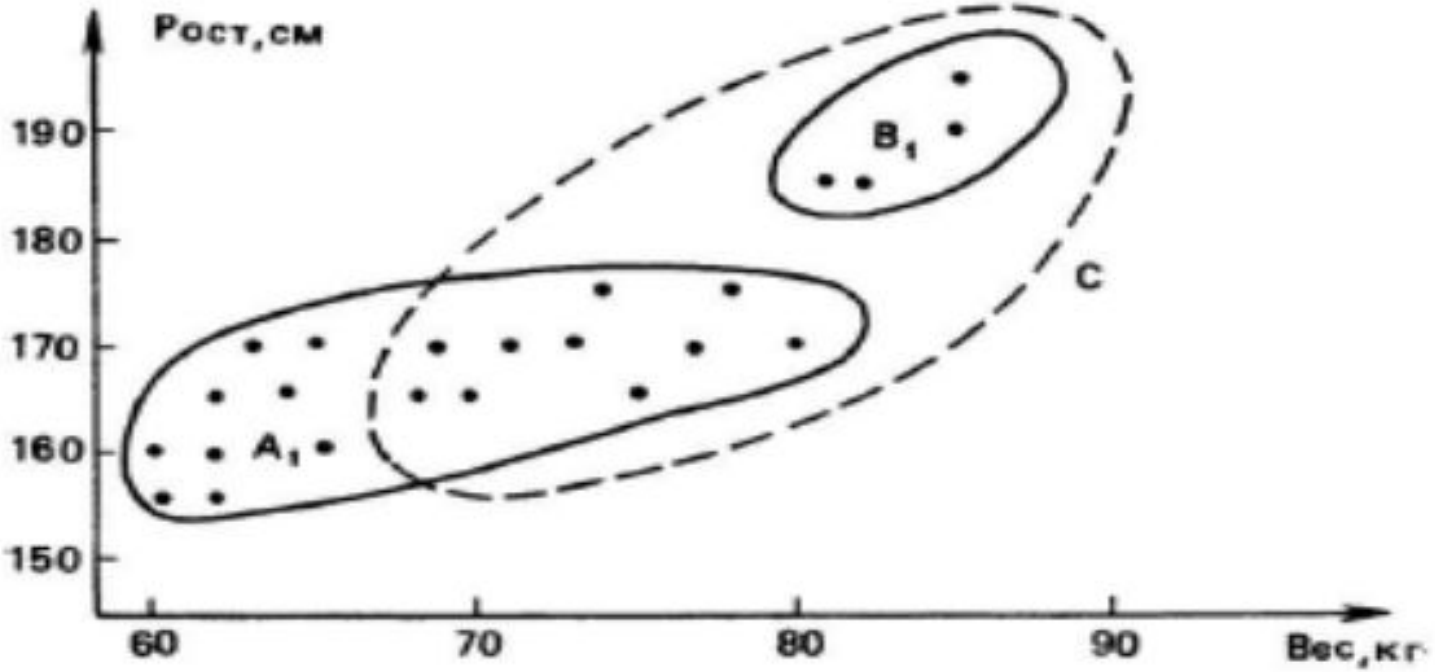
Радиус кластера - максимальное расстояние точек от центра кластера.

При неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками. Существуют различные способы нормирования исходных данных, например: деление исходных данных на среднеквадратичное отклонение соотве









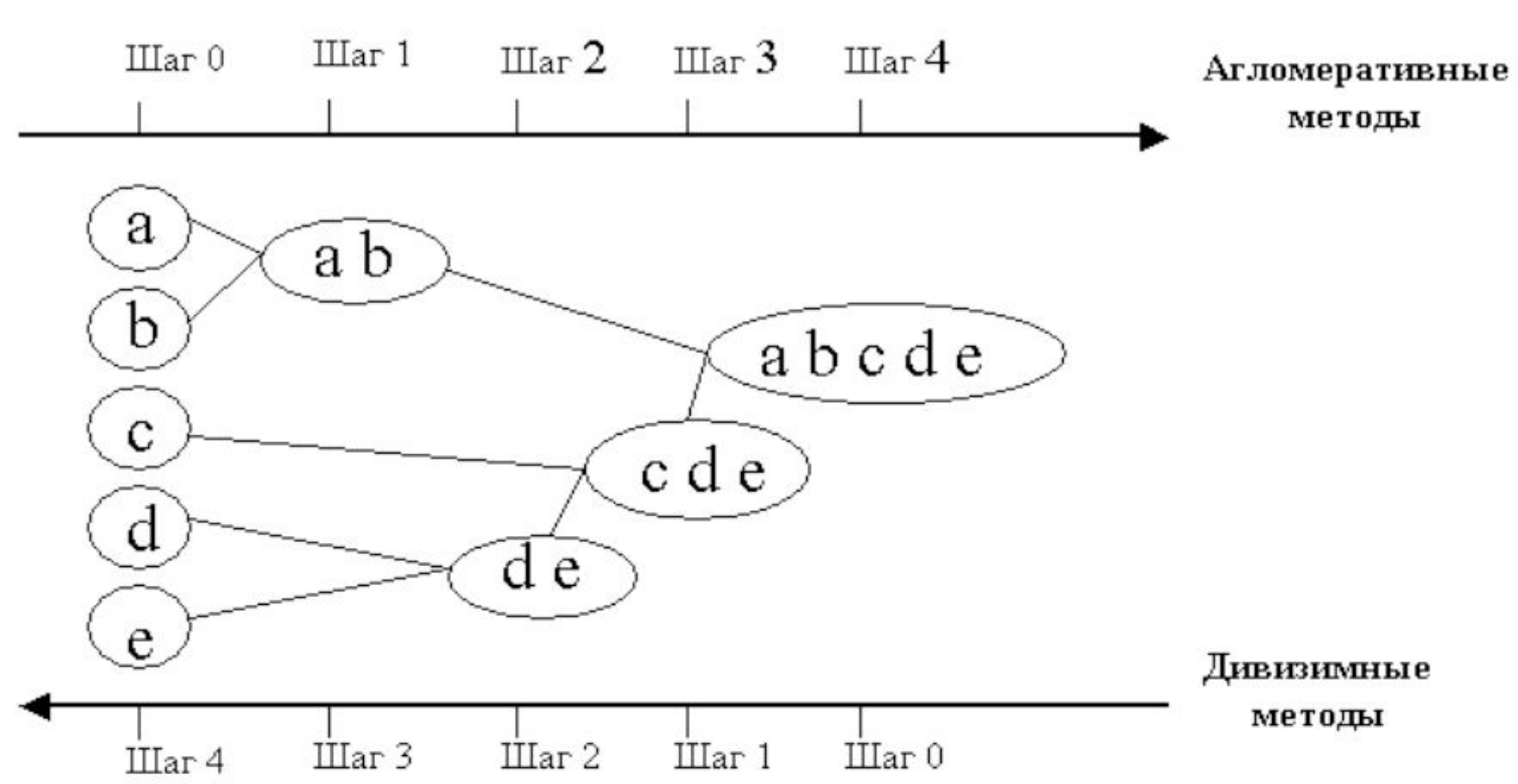
Результат зависит от нормировки признаков

Расстояние между кластерами

- Метод ближнего соседа. Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах.
- Метод наиболее удаленных соседей или полная связь.
- Метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения. Этот метод направлен на объединение близко расположенных кластеров и "стремится" создавать кластеры малого размера.
- Метод невзвешенного попарного арифметического среднего. В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них.

Методы кластеризации: иерархические

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



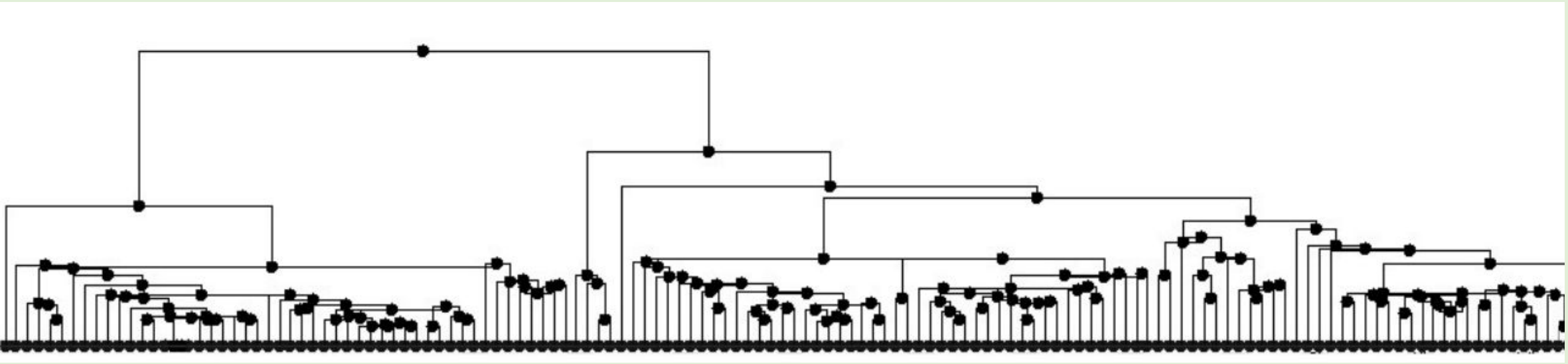
- Иерархические методы кластерного анализа используются при небольших объемах наборов данных.
- Преимуществом иерархических методов кластеризации является их наглядность.
- Иерархические алгоритмы связаны с построением дендрограмм, которые являются результатом иерархического кластерного анализа.
- Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.
- Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Ирисы • Быстрый просмотр

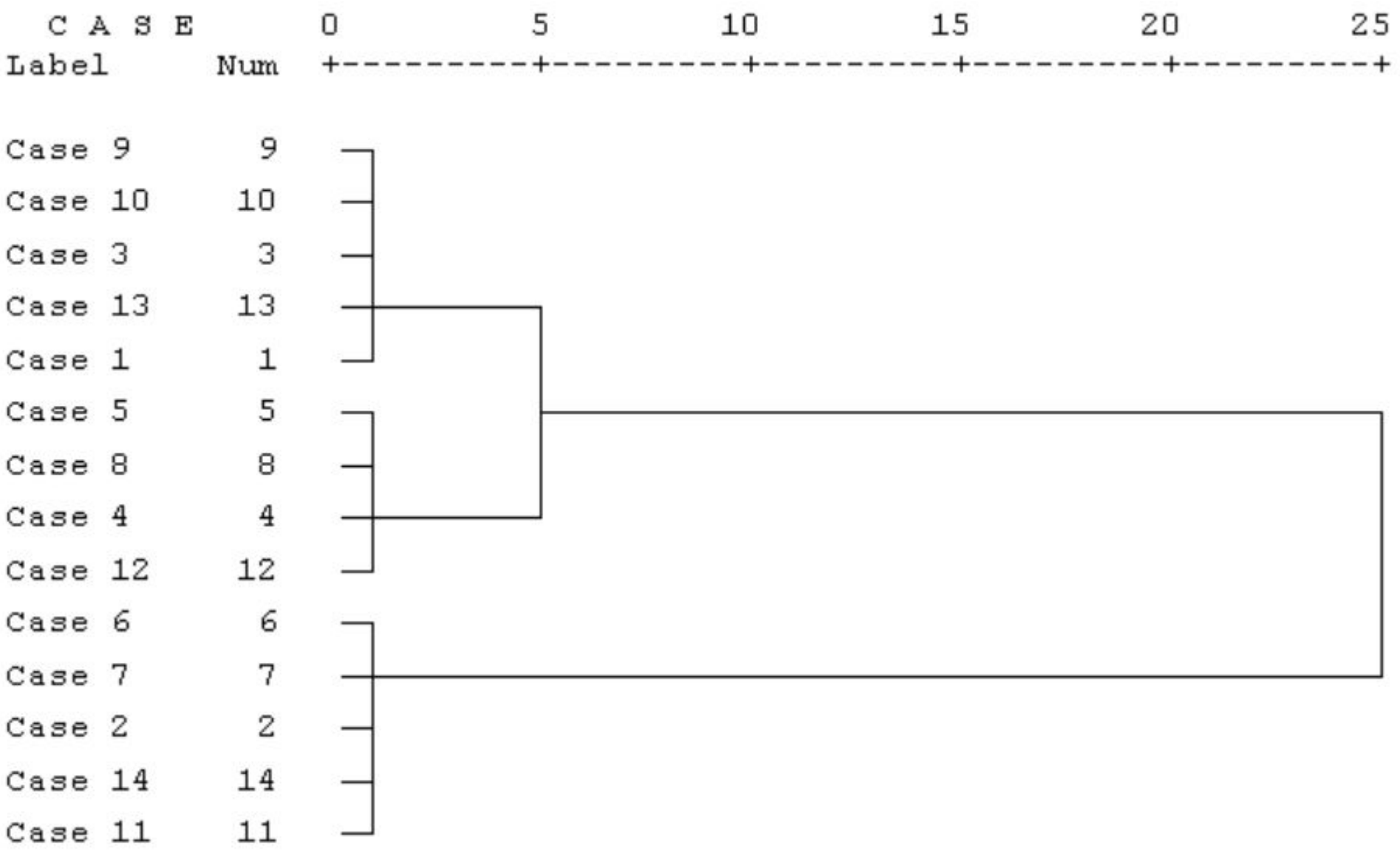
Набор данных

#	Длина чашелисти…	Ширина чашелистика	Длина лепест…	Ширина лепест…	Класс
48	4,60	3,20	1,40	0,20	Iris-setosa
49	5,30	3,70	1,50	0,20	Iris-setosa
50	5,00	3,30	1,40	0,20	Iris-setosa
51	7,00	3,20	4,70	1,40	Iris-versicolor
150	6,40	3,20	4,50	1,50	Iris-versicolor

Таблица Форма



- Существует проблема определения числа кластеров. Иногда можно априорно определить это число. Однако в большинстве случаев число кластеров определяется в процессе агломерации/разделения множества объектов.
- Процессу группировки объектов в иерархическом кластерном анализе соответствует постепенное возрастание коэффициента в протоколе объединения.
- Скачкообразное увеличение значения коэффициента объединения можно определить как переход от сильно связанного к слабо связанному состоянию объектов.
- Оптимальным считается количество кластеров, равное разности количества наблюдений и количества шагов до скачкообразного увеличения коэффициента.

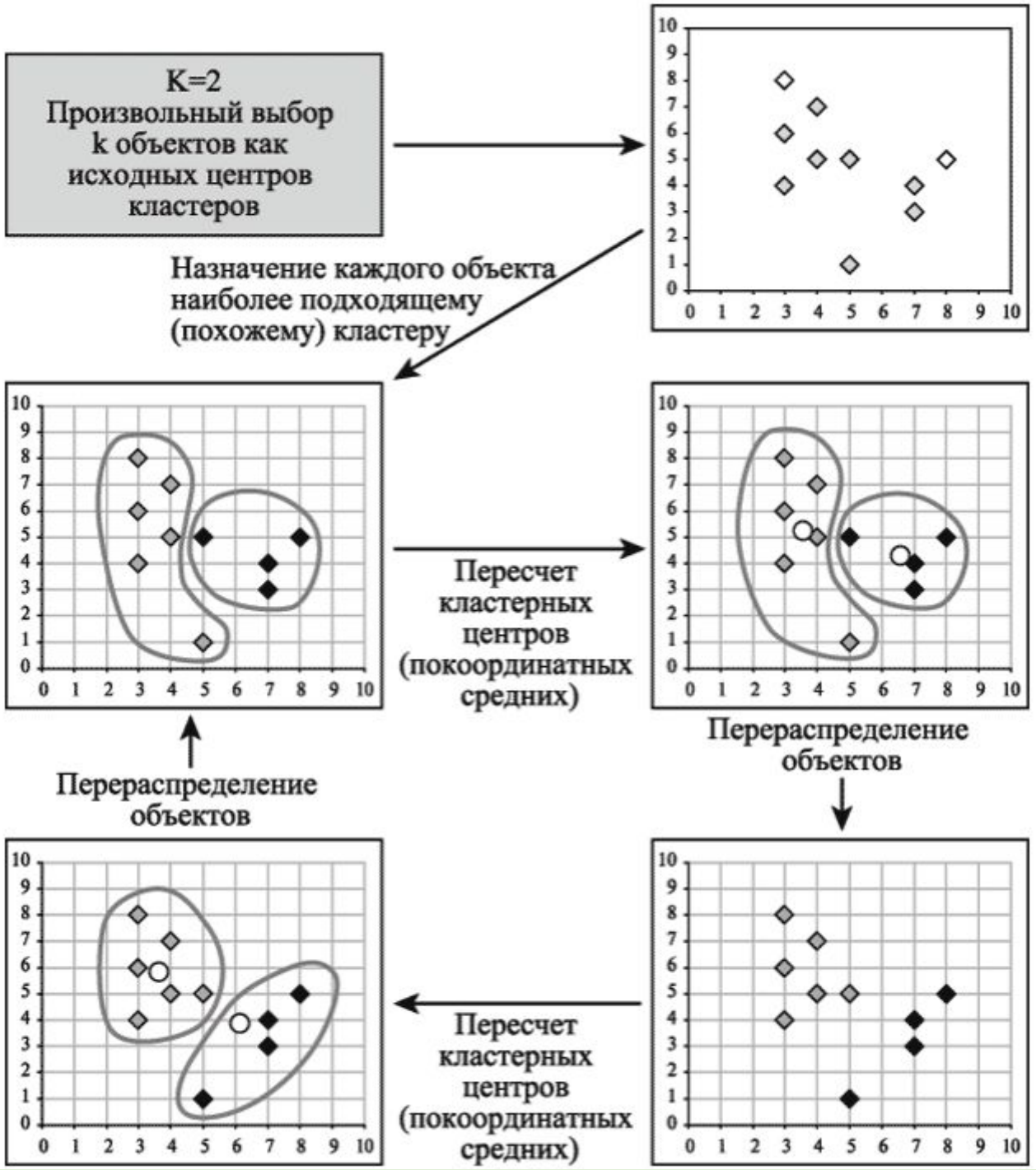

C A S E
Label Num
0
5
10
15
20
25
Case 9 9
Case 10 10
Case 3 3
Case 13 13
Case 1 1
Case 5 5
Case 8 8
Case 4 4
Case 12 12
Case 6 6
Case 7 7
Case 2 2
Case 14 14
Case 11 11

Итеративные методы. Алгоритм k-средних (k-means)

- Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый быстрым кластерным анализом.
- В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для метода k-means необходимо иметь гипотезу о наиболее вероятном количестве кластеров.
- Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга.
- Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

K-средних

- Выбор начальных центроидов может осуществляться следующим образом:
  - выбор k-наблюдений для максимизации начального расстояния;
  - случайный выбор k-наблюдений;
  - выбор первых k-наблюдений.
- В результате каждый объект назначен определенному кластеру.
- Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.
- Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:
  - кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;



Достоинства алгоритма k-средних:

- простота использования;
- быстрота использования;
- понятность и прозрачность.

Недостатки алгоритма k-средних:

- алгоритм чувствителен к выбросам, искажающим среднее.

Для решения этой проблемы - модификация алгоритм k-медианы;

- алгоритм может медленно работать на больших базах данных.

Возможным решением данной проблемы является

EM 15

Проверка качества кластеризации

- После получений результатов кластерного анализа методом k-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга).
- Для этого рассчитываются средние значения для каждого кластера.
- При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

Алгоритм g-means

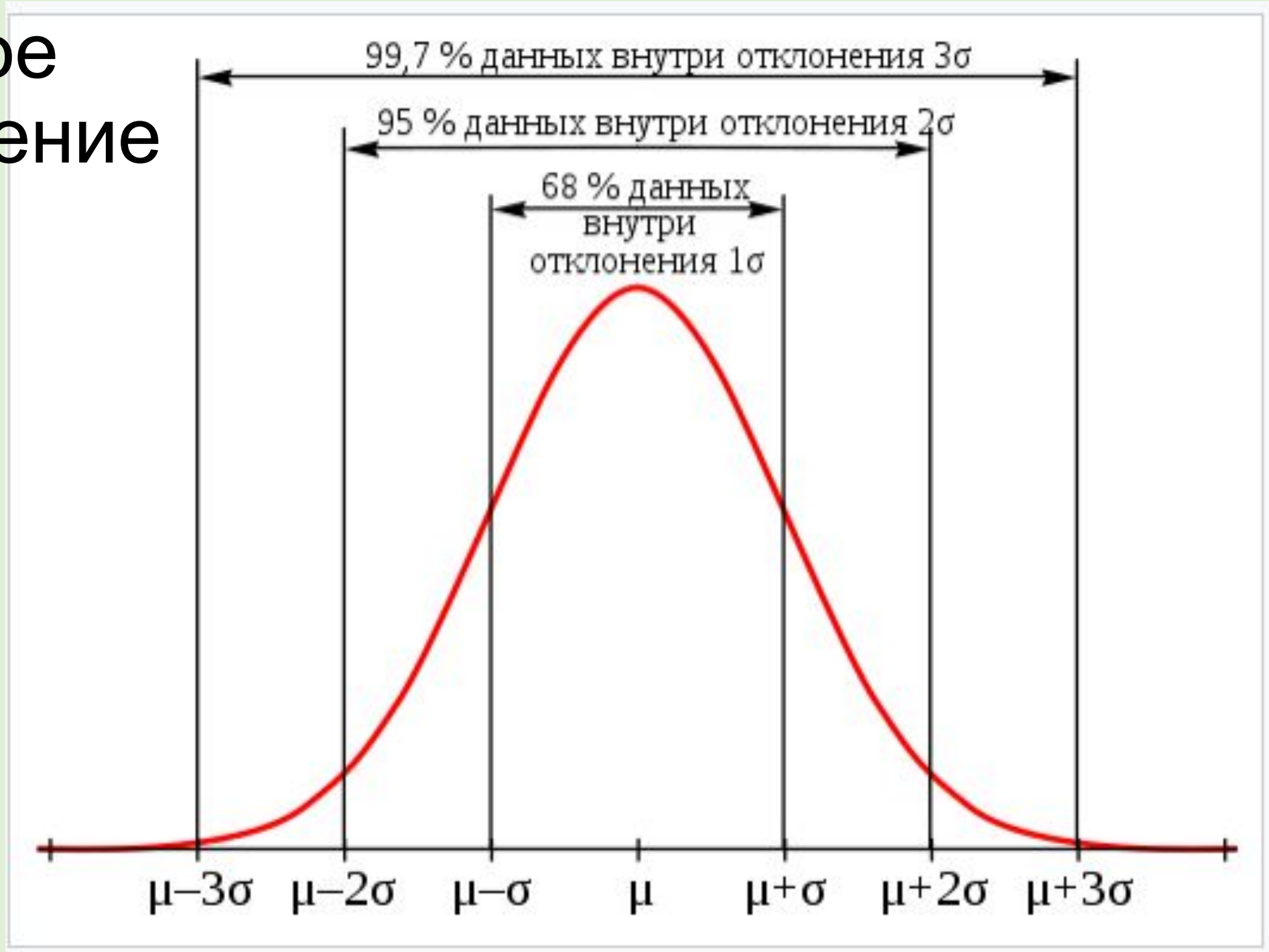
- Недостаток k-means - отсутствие ясного критерия для выбора оптимального числа кластеров.
- G-means - популярный алгоритм кластеризации с автоматическим выбором числа кластеров.
- Предположение - обрабатываемые данные подчиняются **распределению Гаусса** (нормальному **распределению**).
- Алгоритм итеративный: на каждом шаге с помощью k-means строится модель с определенным числом кластеров.

Обычно g-means начинает работу с небольшого значения k=1.

- На каждой итерации увеличение k производится за счет разбиения кластеров, в которых данные не соответствуют гауссовскому распределению.

Нормальное распределение

Для нормального распределения количество значений, отличающиеся от среднего на число, меньшее чем одно стандартное отклонение, составляют 68,27 % выборок.



Ирисы Фишера

Таблица 7.1

№	Длина чашелистника	Ширина чашелистника	Длина лепестка	Ширина лепестка	Класс
1	5,1	3,5	1,4	0,2	Iris setosa
2	4,9	3,0	1,4	0,2	Iris setosa
3	4,7	3,2	1,3	0,2	Iris setosa
4	4,6	3,1	1,5	0,2	Iris setosa
5	5,0	3,6	1,4	0,2	Iris setosa
51	7,0	3,2	4,7	1,4	Iris versicolor
52	6,4	3,2	4,5	1,5	Iris versicolor
53	6,9	3,1	4,9	1,5	Iris versicolor
54	5,5	2,3	4,0	1,3	Iris versicolor
55	6,5	2,8	4,6	1,5	Iris versicolor
101	6,3	3,3	6,0	2,5	Iris virginica
102	5,8	2,7	5,1	1,9	Iris virginica
103	7,1	3,0	5,9	2,1	Iris virginica
104	6,3	2,9	5,6	1,8	Iris virginica
105	6,5	3,0	5,8	2,2	Iris virginica

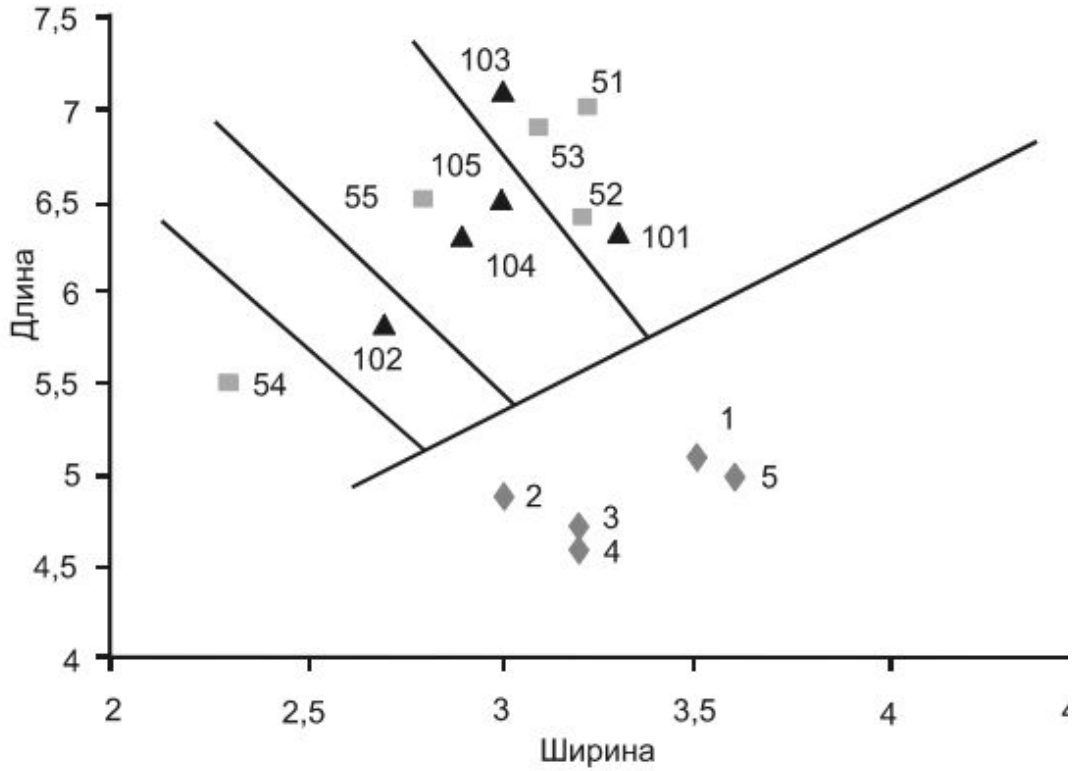


Рис. 7.1. Разделение ирисов на кластеры линиями

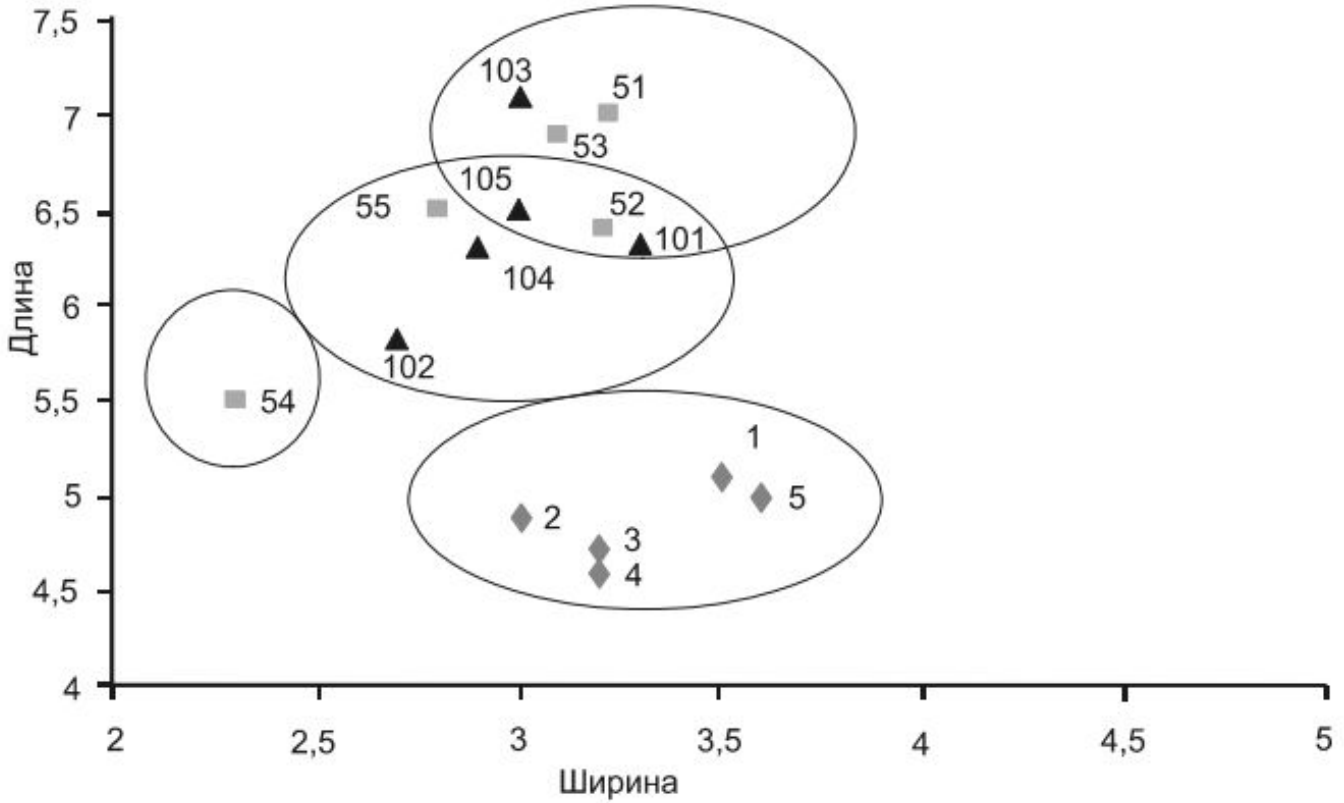


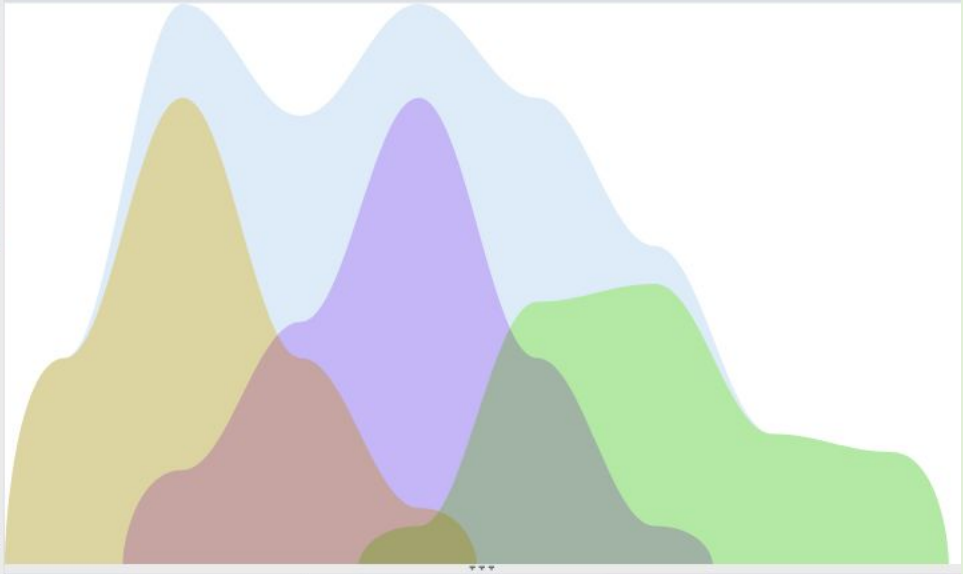
Рис. 7.2. Разделение ирисов на кластеры с использованием Венских диаграмм

Профили кластеров

Пакеты | Кластеризация_Ирисы | Модуль1 | Сценарий | Кластеризация (k-means) | Визуализаторы | Профили кластеров

Фильтрация | Значимость | Сравнение

#	Метка кластера	Поддержка	Длина чашелист…	Ширина чашели…	Длина лепес
1	Итого	100%	66	61	
2	Кластер 2	37%	53	55	
3	Кластер 0	33%	61	57	
4	Кластер 1	29%	69	52	



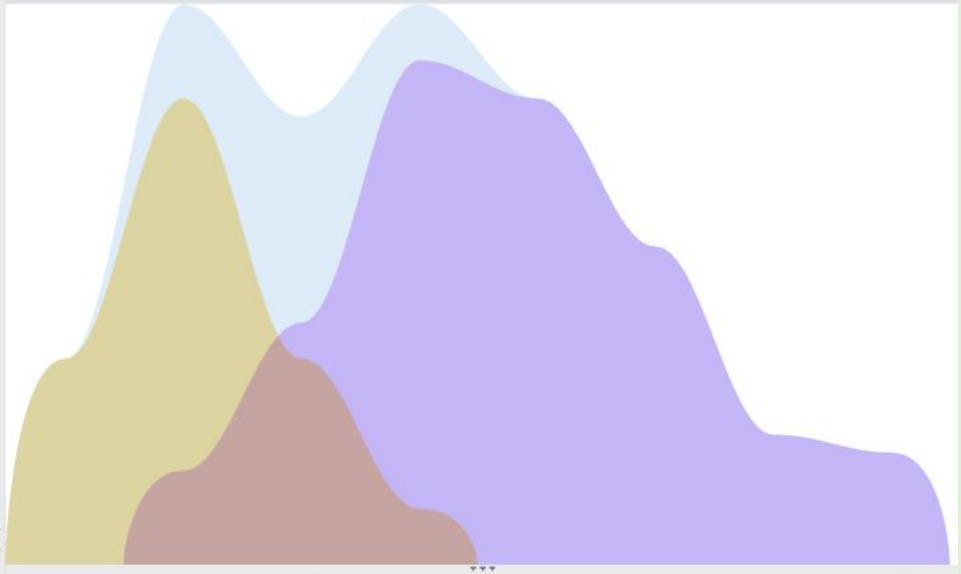
Атрибут	Итого	Кластер 2	Кл
Значимость	66	53	
Минимум	4,30	4,90	
Максимум	7,90	6,70	
Среднее	5,84	5,83	
Сумма	876,50	326,70	
Стандартное от…	0,83	0,42	
Размах	3,60	1,80	
Пропуски	0	0	
Значения	150	56	
Количество ун…			

Профили кластеров

Пакеты | Кластеризация_Ирисы | Модуль1 | Сценарий | Кластеризация (g-means) | Визуализаторы | Профили кластеров

Фильтрация | Значимость | Сравнение

#	Метка кластера	Поддержка	Длина чашелист…	Ширина чашели…	Длина лепес
1	Итого	100%	63	52	
2	Кластер 0	67%	32	29	
3	Кластер 1	33%	58	53	



Атрибут	Итого	Кластер 0	К
Значимость	63	32	
Минимум	4,30	4,90	
Максимум	7,90	7,90	
Среднее	5,84	6,26	
Сумма	876,50	626,20	
Стандартное от…	0,83	0,66	
Размах	3,60	3,00	
Пропуски	0	0	
Значения	150	100	
Количество ун…			

Максимальное ожидание EM-алгоритм

- EM (Expectation maximization - максимальное правдоподобие) – популярный алгоритм кластеризации.
- В основе идеи EM-алгоритма лежит предположение, что любое наблюдение принадлежит ко всем кластерам, но с разной вероятностью. Формируются два дополнительных столбца: *Номер кластера* и *Вероятность принадлежности*.
- Объект должен быть отнесен к тому кластеру, для которого вероятность принадлежности выше.

EM Кластеризация • Разбиение на кластеры • Быстрый просмотр данных

Таблица | Форма

#	12 Номер кластера	9.0 Вероятность принадлежности	9.0 Длина …	9.0 Ширин…	9.0 Длина л…	9.0 Ширина лепест…	ab Класс
69	2	0,83	6,20	2,20	4,50	1,50	Iris-versicolor
70	2	1,00	5,60	2,50	3,90	1,10	Iris-versicolor
71	0	0,91	5,90	3,20	4,80	1,80	Iris-versicolor
72	2	1,00	6,10	2,80	4,00	1,30	Iris-versicolor
73	2	0,70	6,30	2,50	4,90	1,50	Iris-versicolor
74	2	0,98	6,10	2,80	4,70	1,20	Iris-versicolor
75	2	1,00	6,40	2,90	4,30	1,30	Iris-versicolor

- Главным достоинством EM-алгоритма является простота исполнения.

Алгоритм может оптимизировать не только параметры модели, но и делать предположения относительно значений отсутствующих данных.

Это делает EM отличным методом для кластеризации и создания моделей с параметрами. Зная кластеры и параметры модели можно предполагать, что содержит кластер и куда стоит отнести новые данные.

- EM-алгоритм имеет свои недостатки:

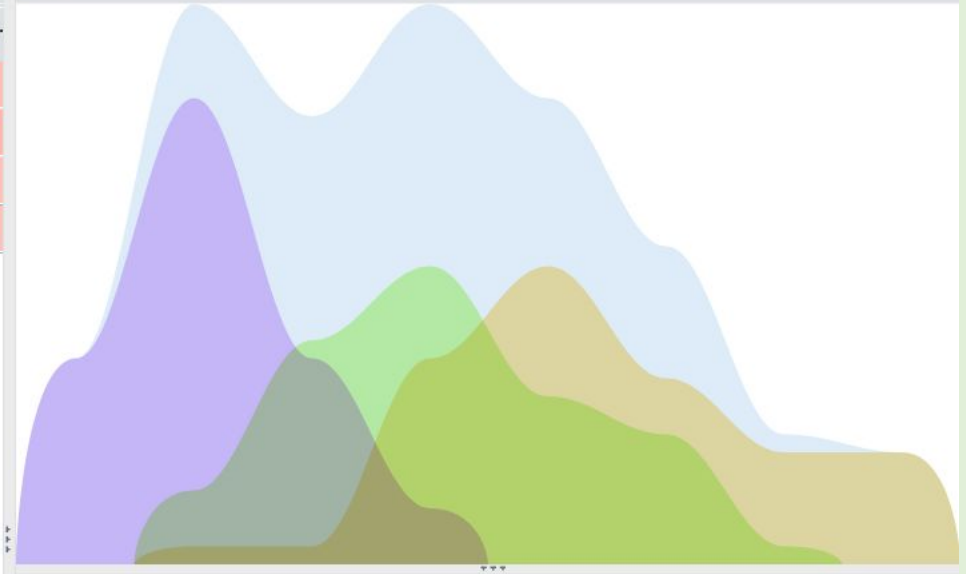
1. С ростом количества итераций падает производительность алгоритма.
2. EM не всегда находит оптимальные параметры и может застрять в локальном оптимуме, так и не найдя глобальный.

Профили кластеров

Пакеты › Кластеризация_Ирисы › Модуль1 › Сценарий › EM Кластеризация › Визуализаторы › Профили кластеров

Фильтрация | Среднее | Сравнение

#	Метка кластера	Поддержка	Длина чашелист…	Ширина чашели…	Длина лепес
1	Кластер 2	33%	5,01	3,42	
2	Итого	100%	5,84	3,05	
3	Кластер 0	33%	5,94	2,76	
4	Кластер 1	34%	6,57	2,98	



Атрибут	Итого	Кластер 2	К
Значимость	64	59	
Минимум	4,30	4,30	
Максимум	7,90	5,80	
Среднее	5,84	5,01	
Сумма	876,50	250,30	
Стандартное от…	0,83	0,35	
Размах	3,60	1,50	
Пропуски	0	0	
Значения	150	50	
Количество ун…			

Кластеризации больших массивов категорийных данных

- Алгоритмы, основанные на *парном вычислении расстояний* (k-means и аналоги) эффективны в основном на числовых данных. Их производительность на массивах записей с большим количеством нечисловых факторов неудовлетворительная.
- На каждой итерации алгоритма требуется попарно сравнивать объекты между собой, а итераций может быть очень много. Для таблиц с миллионами записей и тысячами полей это неприменимо.

Требований к алгоритмам кластеризации номинальных (качественных) данных

- Минимально возможное количество «сканирований» таблицы базы данных;
- Работа в ограниченном объеме оперативной памяти компьютера;
- Работу алгоритма можно прервать с сохранением промежуточных результатов, чтобы продолжить вычисления позже;
- Алгоритм должен работать, когда объекты из базы данных могут извлекаться только в режиме однонаправленного курсора (т.е. в режиме навигации по записям).

Алгоритм CLOPE (кластеризация с наклоном)

- CLOPE (**C**lustering with s**LOPE**) предложен в 2002 году группой китайских ученых.
- Транзакция - некоторый произвольный набор объектов.
- Задача кластеризации - получение такого разбиения всего множества транзакций, чтобы похожие транзакции оказались в одном кластере, а отличающиеся друг от друга — в разных кластерах.
- В основе алгоритма идея максимизации глобальной функции стоимости, которая повышает близость транзакций в кластерах при помощи увеличения параметра **кластерной гистограммы**.

Математическое описание алгоритма

Каждый элемент C_i называется *кластером*, n — **количество транзакций**, A — **длина транзакции**, k — **число кластеров**.

При этом задан параметр r - **коэффициент отталкивания**.

Характеристики кластера C_j :

$D(C_j)$ — множество уникальных объектов в кластере C_j;

$Occ(i, C_j)$ — количество вхождений объекта i в кластер C_j ;

$S(C_j) = \sum_{i \in D(C_j)} Occ(i, C_j) = \sum_{t_i \in C_j} |\ t_i|$ — общее количество объектов в кластере C_j;

$W(C_j) = |\ D(C_j)\ |$ — **ширина** кластера;

$H(C_j) = S(C_j) \ / \ W(C_j)$ — **высота** кластера.

При использовании метода CLOPE количество кластеров подбирается автоматически и зависит от коэффициента отталкивания — параметра, определяющего уровень сходства транзакций внутри кластера. Коэффициент отталкивания задается пользователем: чем больше данный параметр, тем ниже уровень сходства транзакций и, как следствие, большее количество кластеров будет создано.

Формула для вычисления глобального критерия — **функции стоимости** текущей кластеризации $C = \{C_1, \dots C_k\}$, обозначаемой, как $Profit(C)$:

$$Profit(C) = \frac{\sum_{i=1}^{k} H(C_i) \times |C_i|}{\sum_{i=1}^{k} |C_i|} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times |C_i|}{\sum_{i=1}^{k} |C_i|}$$

Для построения начального распределения транзакций по кластерам необходимо сделать первый проход по таблице (фаза инициализации). Во время нее для каждой транзакции выбирается кластер таким образом, чтобы функция $Profit(C)$ была максимальной.

Далее происходят "уточняющие" проходы (фаза итерации), где для каждой транзакции производится попытка перемещения в другой кластер так, чтобы $Profit(C) \longrightarrow max$.

Если больше не существует перемещений транзакций, увеличивающих глобальную стоимость $Profit(C)$, то действие алгоритма заканчивается.

Разность между стоимостью кластеризации $Profit(C)$ после перемещения транзакции t_n с кластера C_i на кластер C_j и стоимостью кластеризации непосредственно до этого перемещения называется **ценой перемещения** транзакции. Цена перемещения t_n с кластера C_i на кластер C_j складывается из двух величин:

- $Cost^-(C_i, t_n) = \frac{(S(C_i) - lentgh(t_n)) * (|C_i| - 1)}{W(C_i \backslash t_n)^r} - \frac{S(C_i) * |C_i|}{W(C_i)^r}$ — **цена удаления** транзакции t_n из кластера C_i.

- $Cost^+(C_j, t_n) = \frac{(S(C_j) + lentgh(t_n)) * (|C_j| + 1)}{W(C_j \cup t_n)^r} - \frac{S(C_j) * |C_j|}{W(C_j)^r}$ — **цена добавления** транзакции t_n на кластер C_j.

ЗАКЛЮЧЕНИЕ В этой статье предлагается новый алгоритм категориальной кластеризации данных, называемый CLOPE, основанный на интуитивной идее увеличения отношения высоты к ширине кластерной гистограммы.

Идея обобщается с помощью параметра отталкивания, который контролирует плотность транзакций в кластере и, следовательно, результирующее количество кластеров.

Простая идея, лежащая в основе CLOPE, делает его быстрым, масштабируемым и экономящим память при кластеризации больших разреженных транзакционных баз данных с большими размерами.

Наши эксперименты показывают, что CLOPE довольно эффективен при поиске интересных кластеров, даже если в нем явно не указана какая-либо метрика межкластерного различия.

Более того, CLOPE не очень чувствителен к порядку данных и не требует большого знания предметной области для управления количеством кластеров.

Эти функции делают CLOPE хорошим алгоритмом кластеризации, а также предварительной обработки для интеллектуального анализа транзакционных данных, таких как данные о рыночной корзине и данные об использовании Интернета.

Описательная аналитика и Кластеризация

- Кластеризация транзакционных данных имеет много общего с анализом ассоциаций: выявляют скрытые зависимости в наборах данных.
- Кластеризация дает общий взгляд на совокупность данных, ассоциативный анализ находит конкретные зависимости между атрибутами.
- Ассоциативные правила сразу пригодны для использования, тогда как кластеризация чаще всего используется как первая стадия анализа.