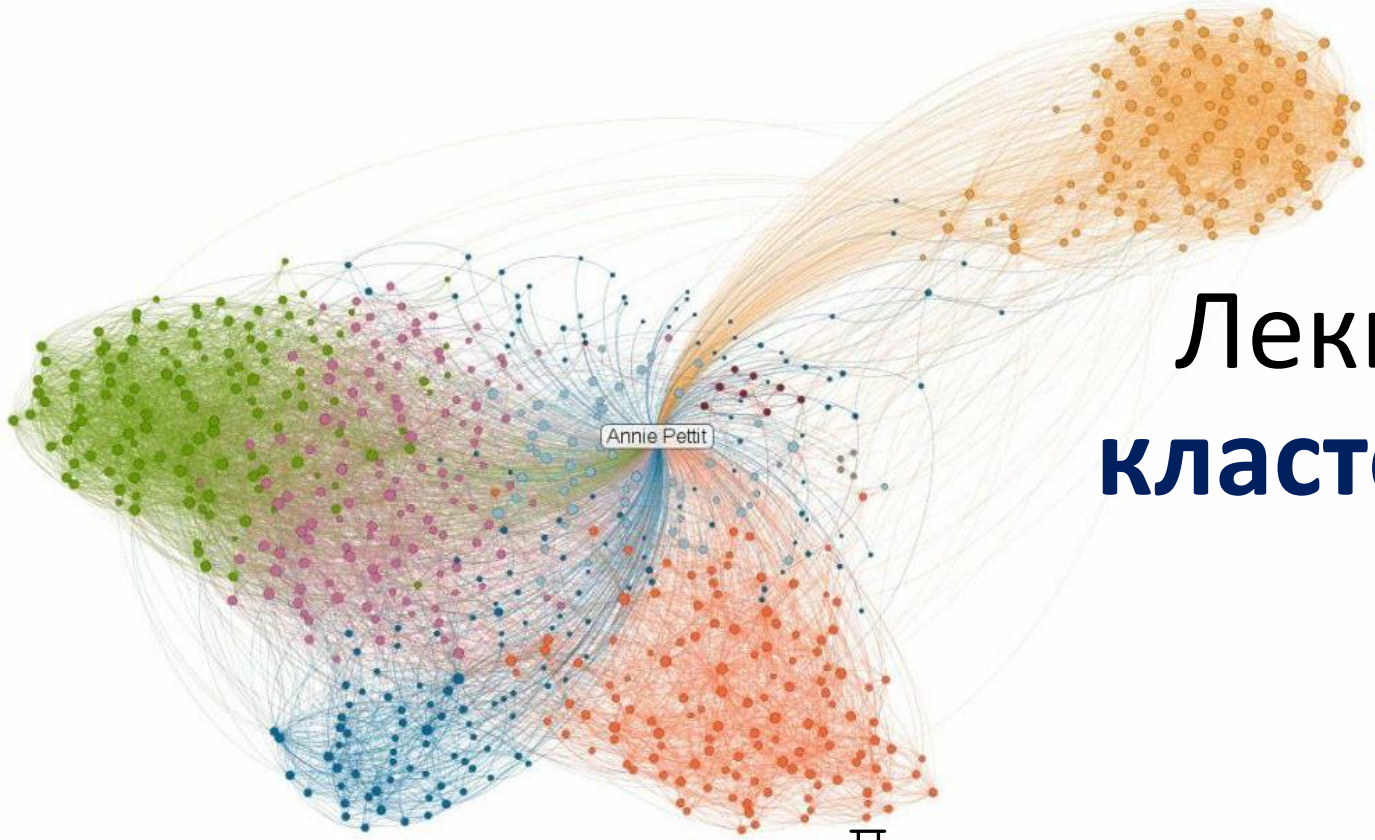


Лекция 4. **Основы кластерного анализа**

Дисциплина: **Интеллектуальный анализ данных, текстов и изображений**

Лектор: к.т.н. **Буров Сергей Александрович**

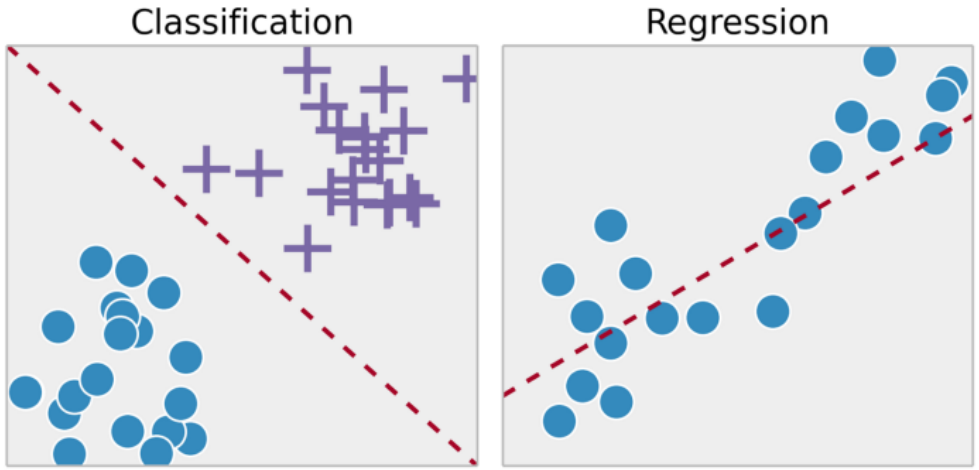
burov-sa@ranepa.ru

Вопросы лекции:

1.Постановка задачи кластерного анализа. Кластер и кластеризация
2. Базовые алгоритмы кластеризации
3. Кластерный анализа на Python

Введение

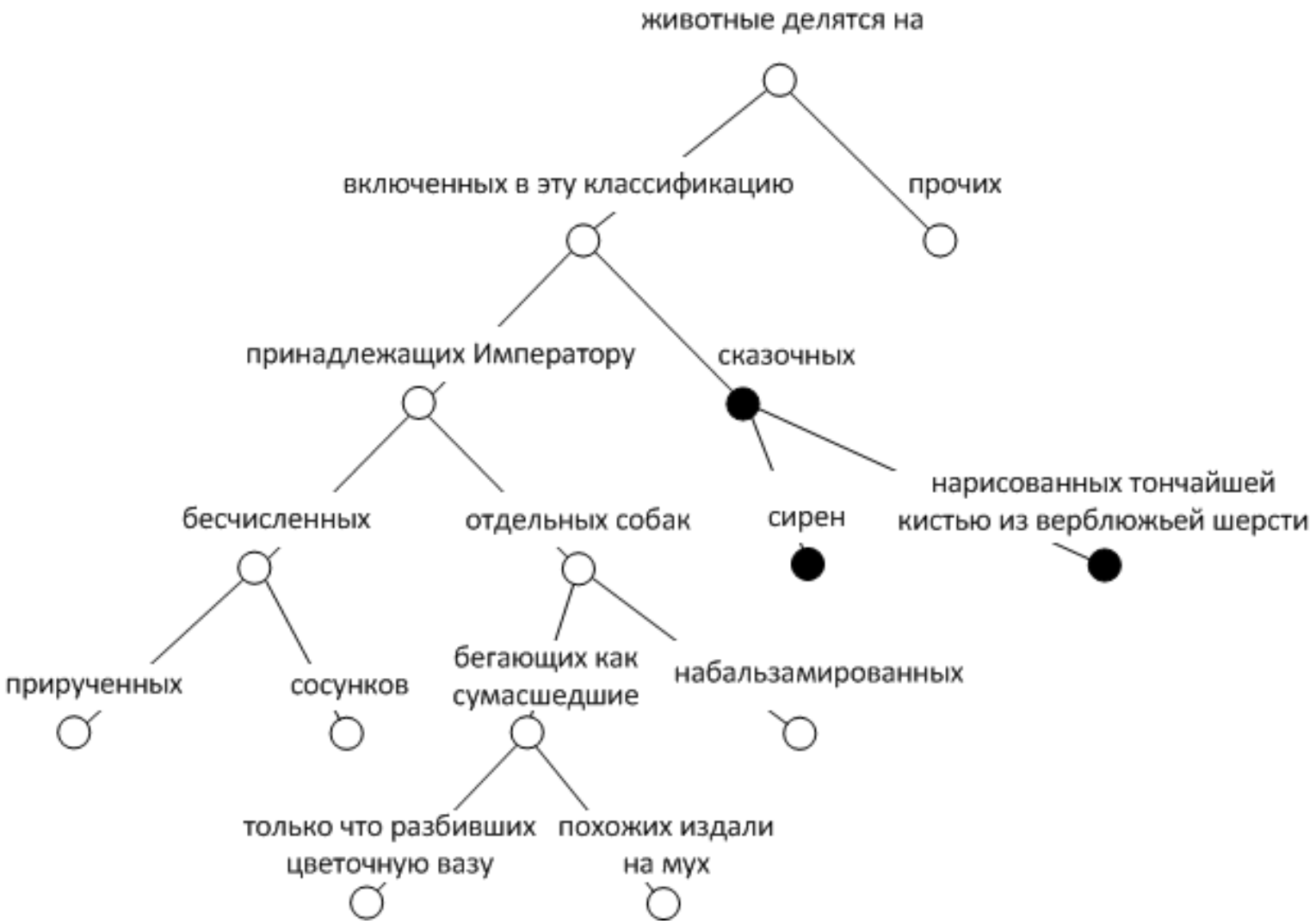

МАШИННОЕ ОБУЧЕНИЕ
Данные заранее категоризированы
или численные
Данные никак не размечены
С учителем
Без учителя
Предсказать категорию
Предсказать значение
Разделить по схожести
Найти зависимости
Выявить последовательности
Классификация
«Разложи носки по цвету»
Регрессия
«Разложи галстуки по длине»
Кластеризация
«Разложи похожие вещи по кучкам»
Ассоциация
«Найди какие вещи я часто ношу вместе»
Уменьшение Размерности (обобщение)
«Собери из вещей лучшие наряды»


Classification
Regression

Введение

Древнекитайская классификация животных

(Хорхе Луис Борхес, *Другие исследования:* 1937—1952).



Введение

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.

Постановка задачи кластерного анализа

Кластерный анализ (`Data clustering`) — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые **кластерами**, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Кластер — группа схожих по своим характеристикам объектов
— несколько однородных элементов, объединенных естественным или искусственным путем в единую структуру

Математическая постановка задачи кластеризации

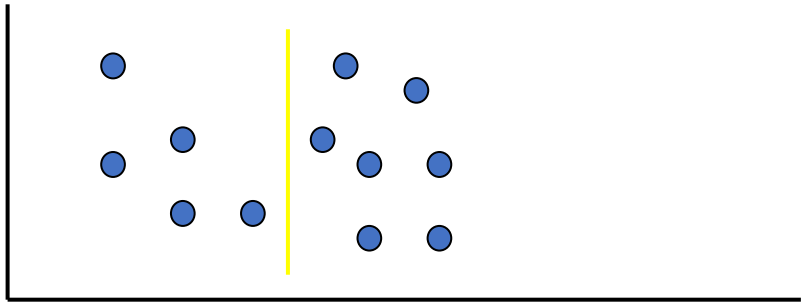
Пусть X — множество объектов, Y — множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $\rho(x,x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1,\dots,x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые *кластерами*, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного *критерия качества* кластеризации.

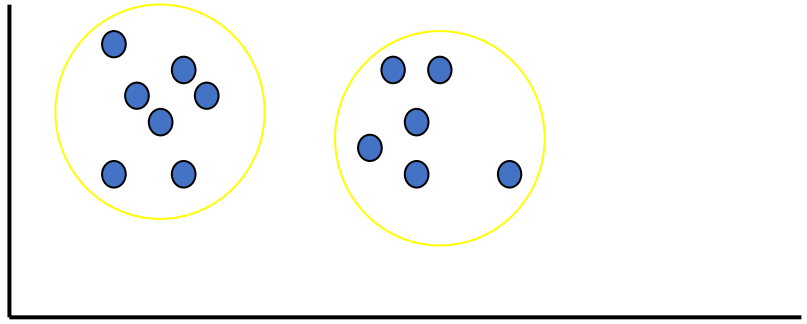
Задачи кластерного анализа

Кластерный анализ (Data clustering) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

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



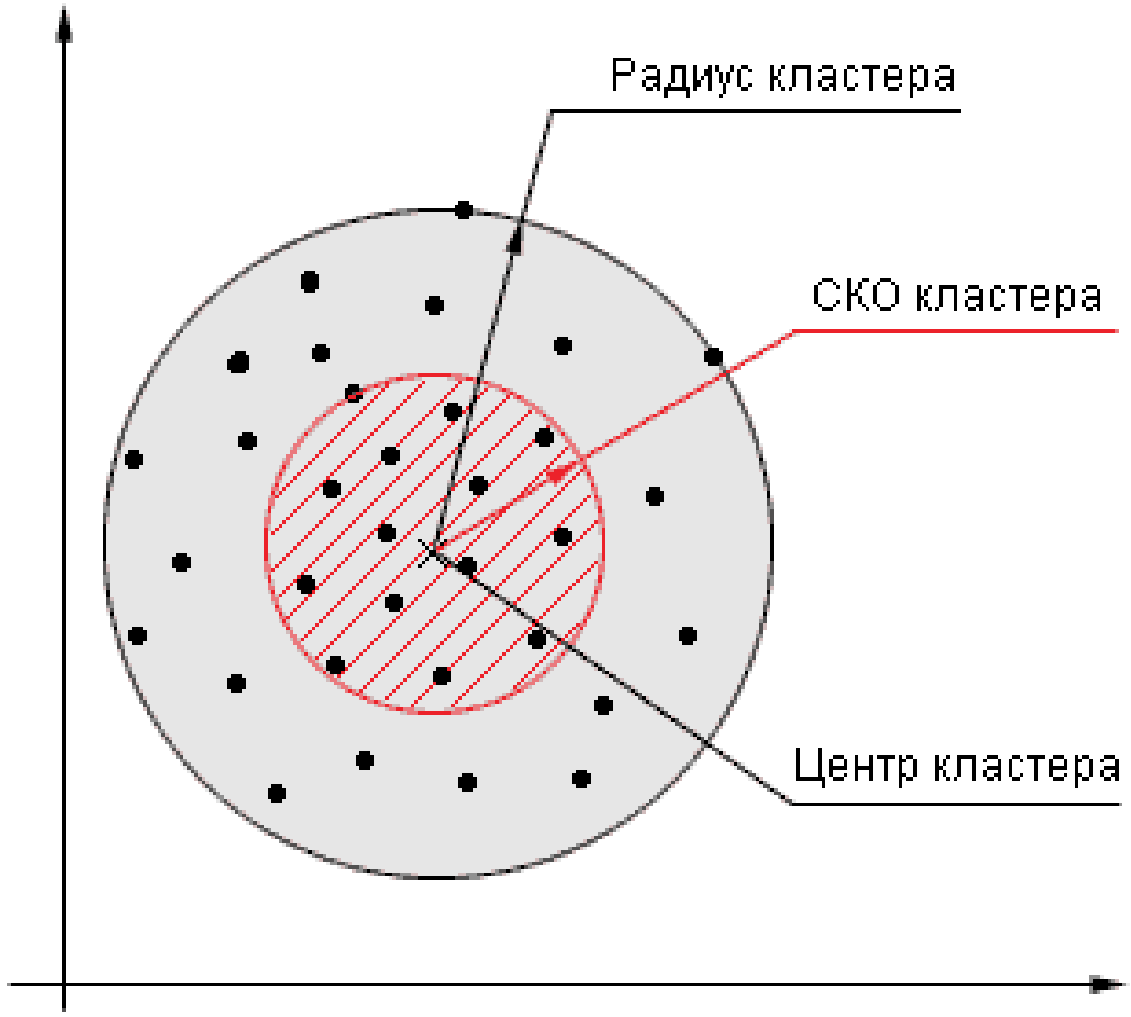
2. **КЛАССИФИКАЦИЯ (естественное расслоение на четко выраженные кластеры). Во второй постановке задача не всегда имеет решение.**



Параметры кластера

Основные:

1) Плотность
2) Дисперсия
3) Размер (радиус)
4) Форма
5) Отделимость



Параметры кластера

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.

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

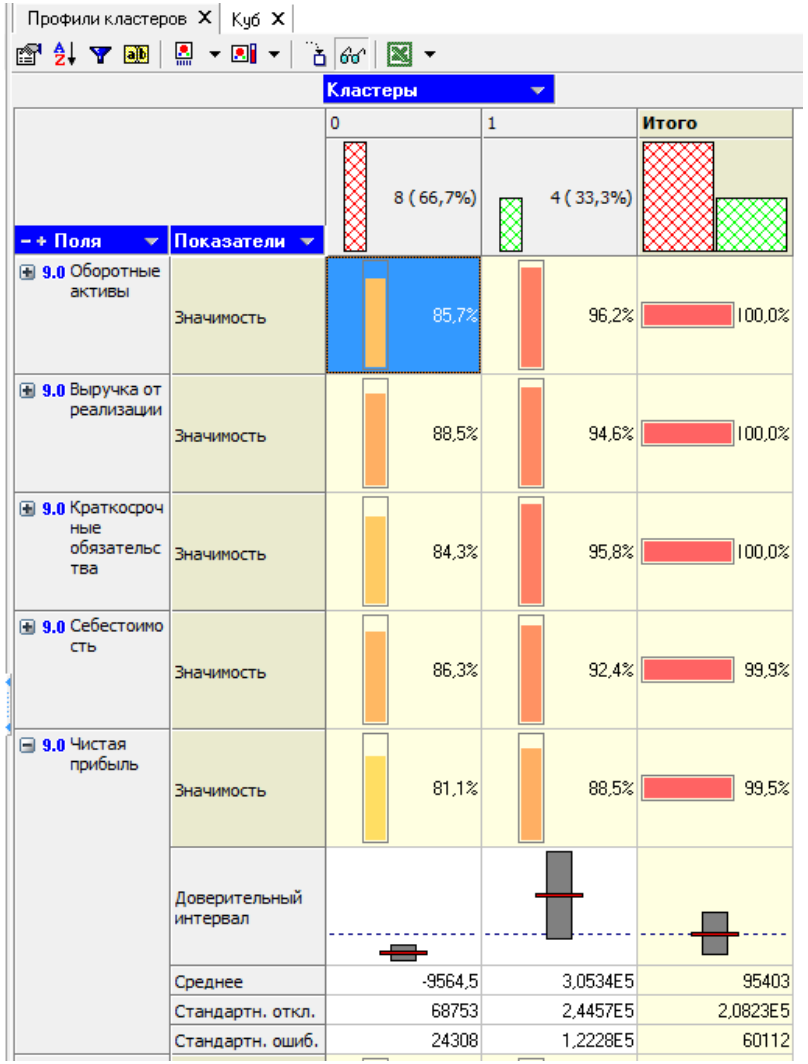
Дисперсия характеризует степень рассеяния точек в пространстве относительно центра кластера, т.е. насколько близко друг к другу расположены точки кластера.

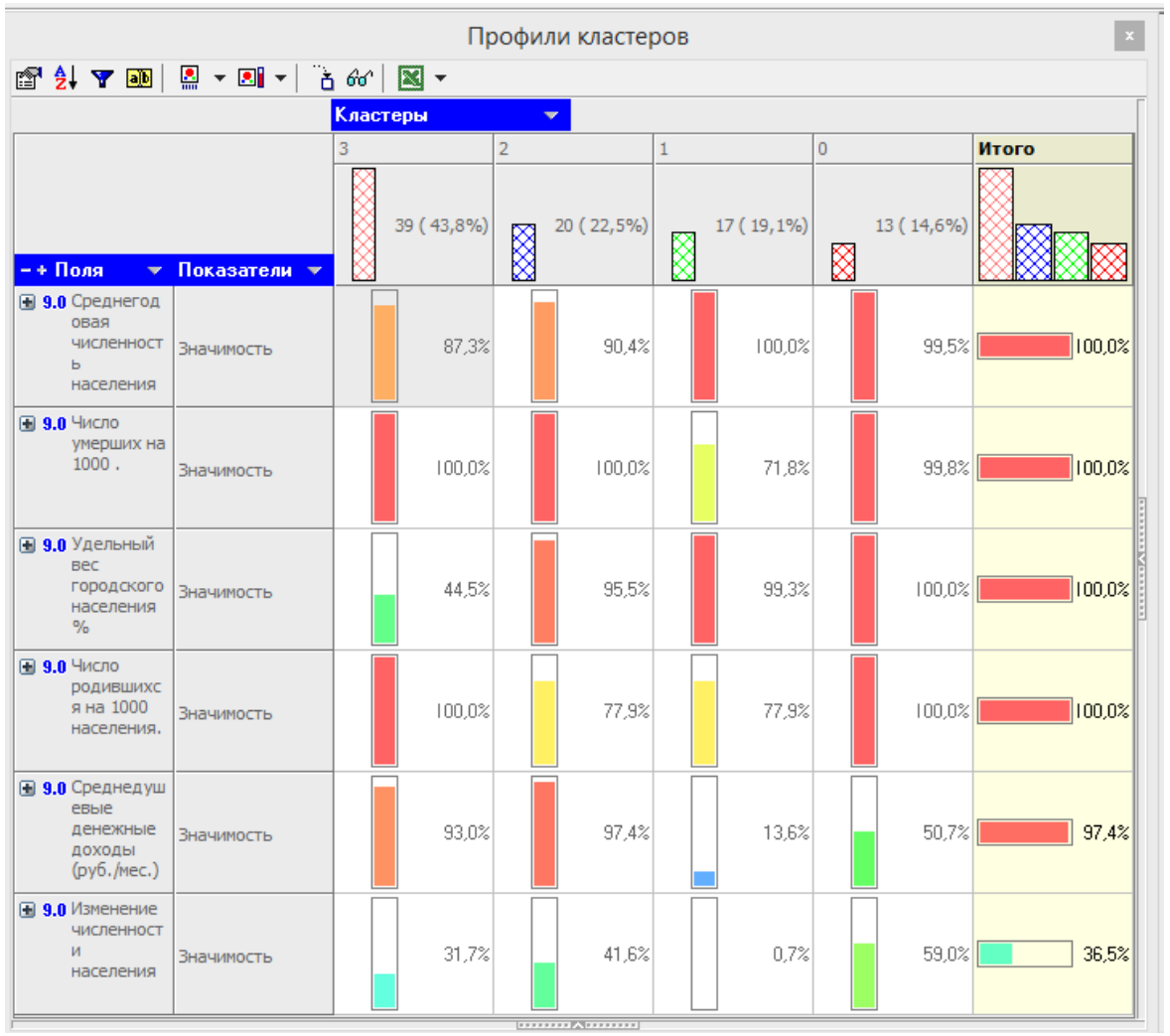
Форма – это расположение точек в пространстве. Если кластеры имеют удлиненную форму, то вместо размера можно вычислить его «связность» - относительную меру расстояния между точками.

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.

Профили кластеров

Профили кластеров предназначен для просмотра различных статистических показателей кластеров, просмотра структуры кластеров и сравнения их между собой





Как измерять расстояние между объектами?

Метрики кластерного анализа

Метрика – способ задания расстояния между объектами

Метрики кластерного анализа

1. Евклидово расстояние *(Euclidean distances)*. Наиболее общий тип расстояния. Хорошо известное из школьного курса как геометрическое расстояние. Вычисляется по формуле (по исходным, а не по стандартизованным данным):

$$dist(x,y) = [\Sigma_i (x_i - y_i)^2]^{1/2}$$

2. Квадрат евклидова расстояния *(Squared Euclidean distances)***.**
Применяется, чтобы придать большие веса более отдаленным друг от друга объектам:

$$dist(x,y) = \Sigma_i \ (xi - yi)2$$

3. Расстояние городских кварталов
*(City-block (Manhattan) distances)***.** В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

$$dist(x,y) = \ \Sigma_i \ |xi - yi|$$

Метрики кластерного анализа

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

$$dist(x,y) = max(|xi - yi|)$$

5. Степенное расстояние. Иногда желают прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это может быть достигнуто с использованием степенного расстояния:

$$dist(x,y) = (\Sigma i|xi - yi|p)1/r$$

где *r* и *p* - параметры, определяемые пользователем. Если оба они равны 2, то это расстояние совпадает с расстоянием Евклида.

6. Процент несогласия *(Percent disagreement)*. Эта мера используется в тех случаях, когда данные являются категориальными.

$$dist(x,y) = (Количество\ xi \neq yi)/ i$$

Кластеризация производится в многомерном пространстве по нескольким переменным, имеющим различный порядок и единицы измерения.

Как сделать переменные равноправными при образовании кластера?

Стандартизация данных

Стандартизация – среднее всех переменных равно 0, дисперсия 1

1. **Вычислим среднее арифметическое и стандартное отклонение каждой из переменных**

2. **Преобразуем каждое значение наблюдения по формуле:**

$$x_{st} = \frac{x_0 - \overline{x}}{\sqrt{\sigma^2}}$$

Нормализация данных

Нормализация данных – предобработка числовых признаков в обучающих выборках данных с целью приведения их к некоторой общей шкале без потери информации о различных диапазонах

1. **Вычислим среднее арифметическое и стандартное отклонение каждой из переменных**

2. **Преобразуем каждое значение наблюдения по формуле:**

$$x_{norm} = \frac{x_0 - x_{min}}{x_{max} - x_{min}}$$

С помощью метрик можно измерить расстояние между всеми парами объектов.

По какому правилу следует производить дальнейшее объединение в кластеры?

Меры близости

1. **Метод ближайшего соседа** *(одиночная связь, Single linkage)*. Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами («ближайшими соседями») в различных кластерах. Это правило похоже на «нанизывание» объектов для формирования кластеров, и результирующие кластеры имеют тенденцию быть представлены длинными «цепочками».
2. **Метод наиболее удаленного соседа** *(полная связь, Complete linkage)***.** Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах.

Меры близости

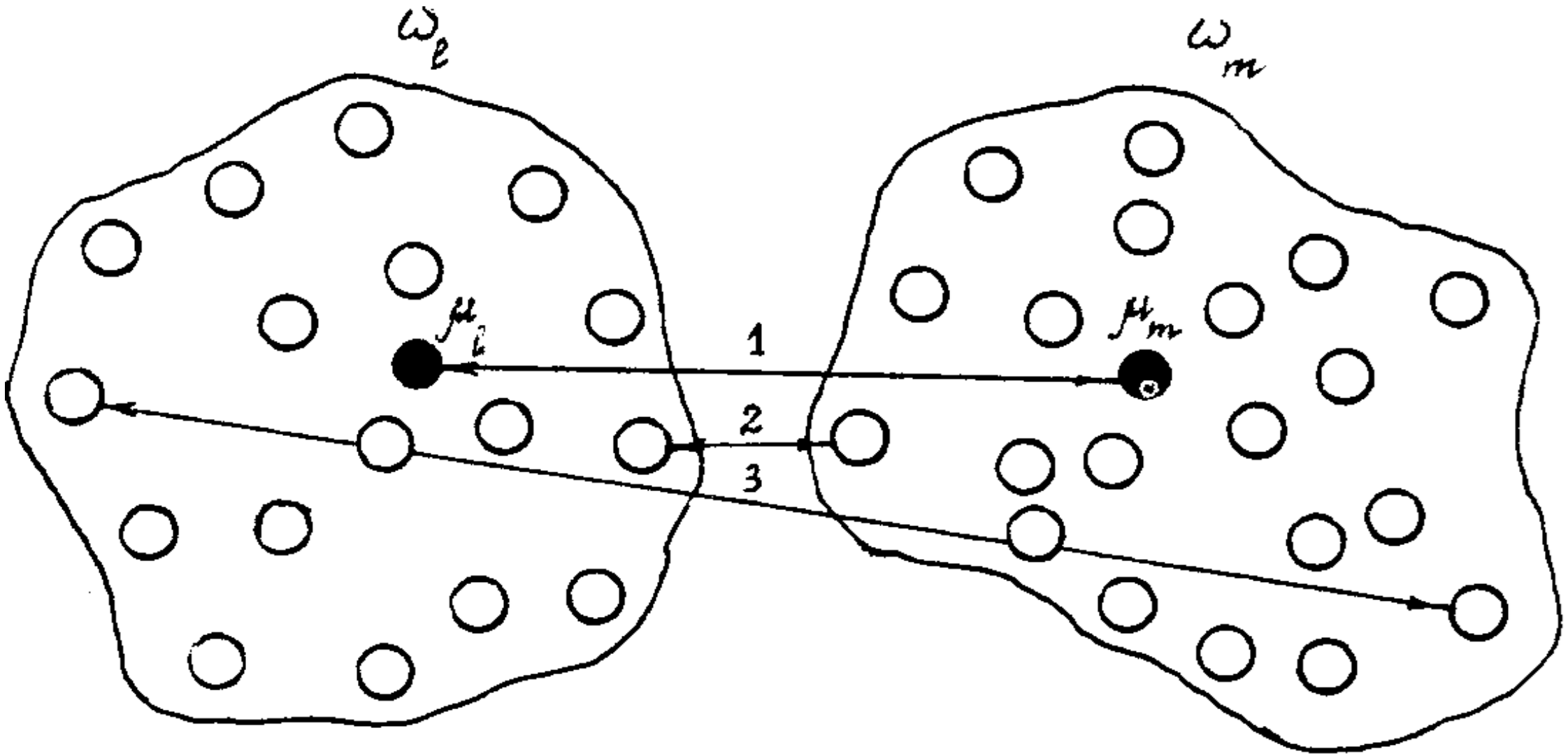
3. Невзвешенное попарное среднее *(Unweighted pair-group average)*. Расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них.

4. Взвешенное попарное среднее *(Weighted pair-group average).* Метод идентичен предыдущему за исключением того, что при вычислениях размер соответствующих кластеров (т. е. число содержащихся в них объектов) используется в качестве весового коэффициента. Поэтому предпочтительней использовать данный метод, если есть предположение о неравных размерах кластеров.

5. Невзвешенный центроидный метод (*Unweighted pair-group centroid)***.** В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

6. Взвешенный центроидный метод *(медиана)***.** Этот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них).

Меры близости



Меры близости

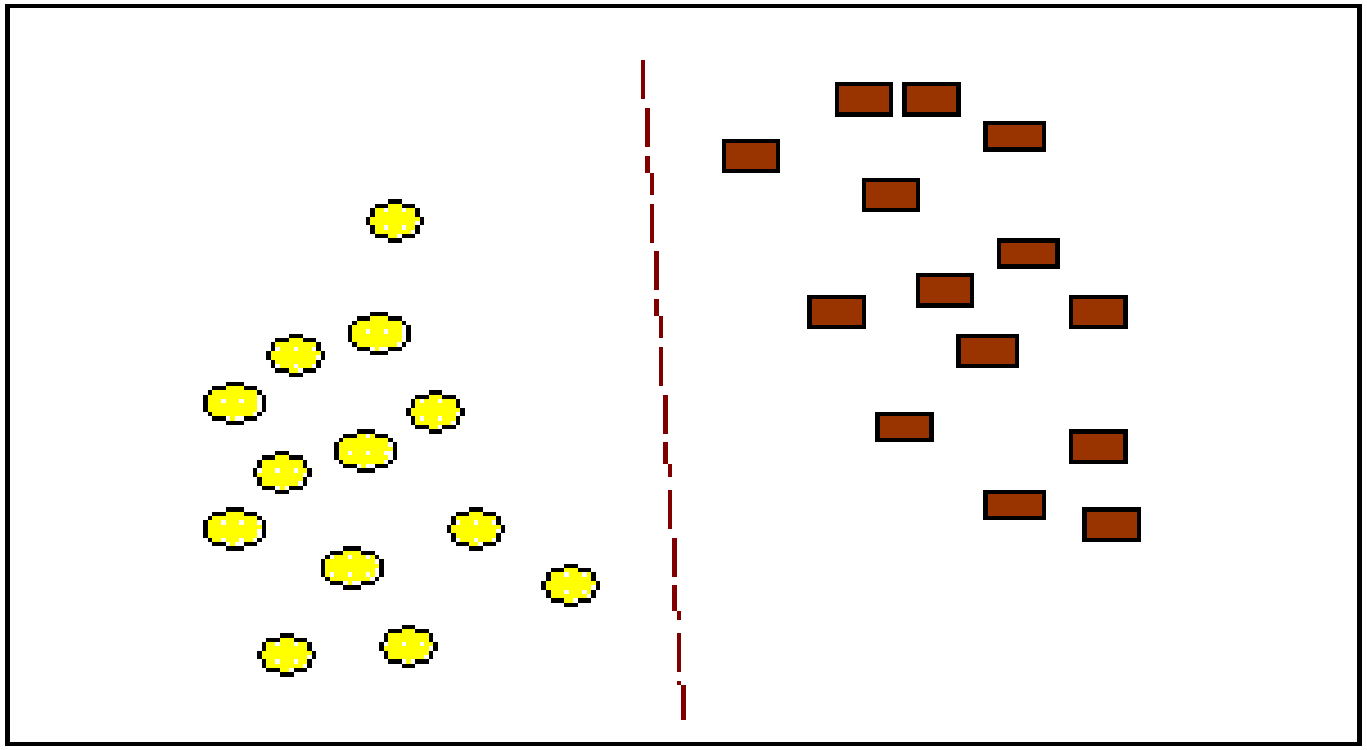
7. Метод Варда (*Ward's method).* Этот метод отличается от всех других методов, поскольку для оценки расстояний между кластерами он использует методы дисперсионного анализа. Метод **минимизирует сумму квадратов** для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

Ошибки суммы квадратов

$$ESS = \sum_i \sum_j \sum_k \left| X_{ijk} - \bar{x}_{i.k} \right|^2$$

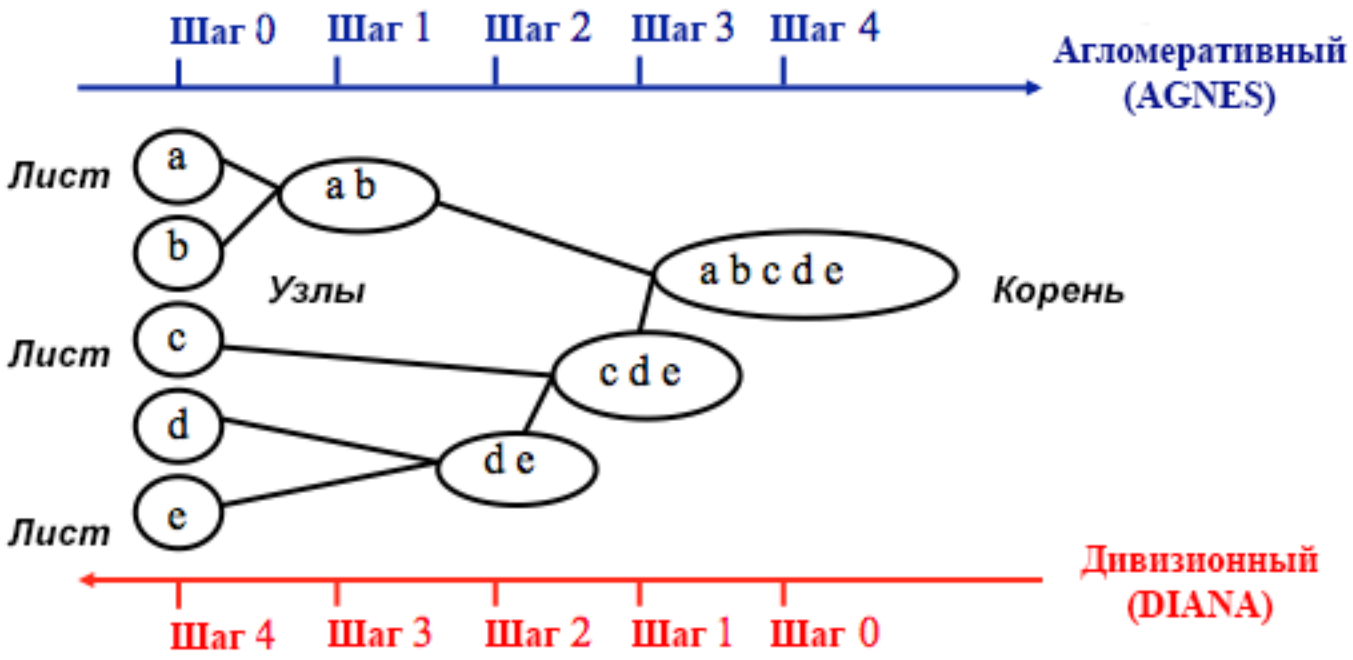
Общая сумма квадратов

$$TSS = \sum_i \sum_j \sum_k \left| X_{ijk} - \bar{x}_{..k} \right|^2$$



Базовые алгоритмы кластеризации

Иерархическая кластеризация



Агломеративные методы (`agglomerative`): новые кластеры создаются путем объединения более мелких кластеров и, таким образом, дерево создается от листьев к стволу;

Дивизивные или дивизионные методы (`divisive`): новые кластеры создаются путем деления более крупных кластеров на более мелкие и, таким образом, дерево создается от ствола к листьям.

Дендограммы

Дендограмма – способ описания и наглядного представления результатов иерархической кластеризации

Метод «к-средних» - k-means

1. Задается число кластеров *k*, которое должно быть сформировано из объектов исходной выборки.

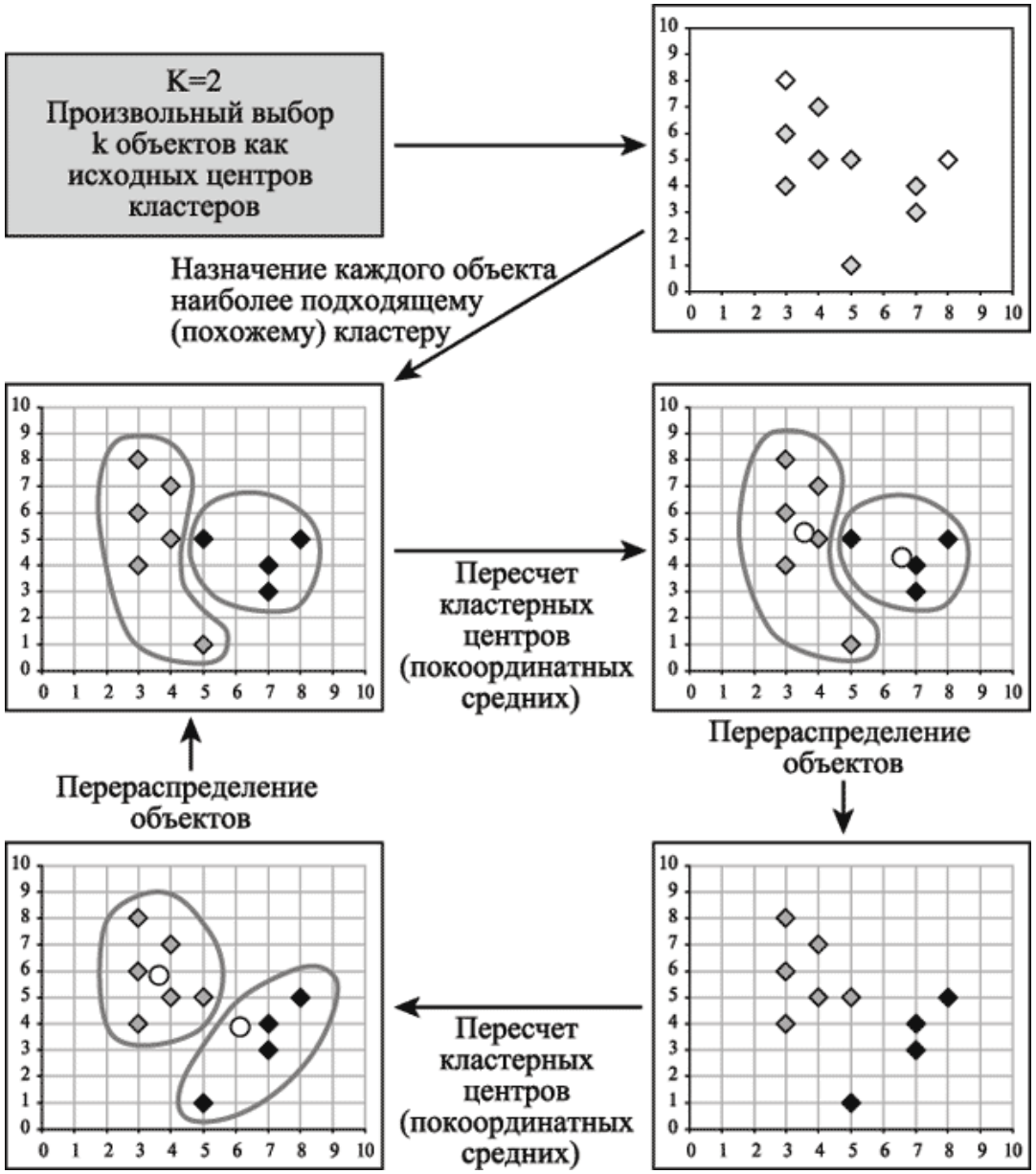
2. Случайным образом выбирается *k* записей исходной выборки, которые будут служить начальными центрами кластеров.

3. Для каждой записи исходной выборки определяется ближайший к ней центр кластера.

4.Производится вычисление **центроидов** – центров тяжести кластеров. Это делается путем определения среднего для значений каждого признака для всех записей в кластере. Затем старый центр кластера смещается в его центроид.

Шаги 3 и 4 повторяются до тех пор, пока центроиды кластеров не перестанут изменяться. В качестве критерия сходимости чаще всего используется критерий суммы квадратов ошибок между центроидом кластера и всеми вошедшими в него записями:

Метод «к-средних» - k-means. Пример k=2



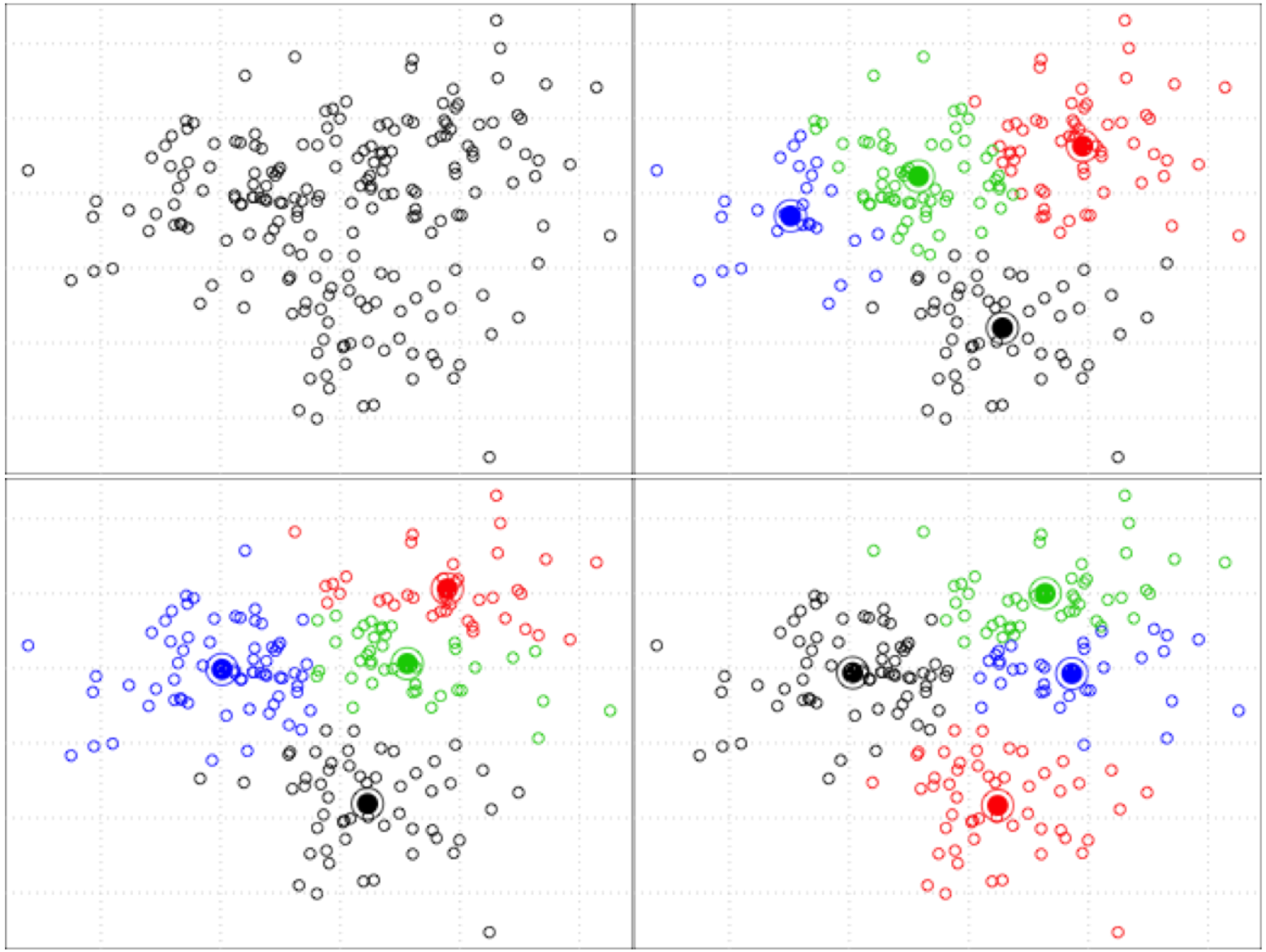
Метод «к-средних» - k-means.

1. Не гарантируется достижение глобального минимума суммарного квадратичного отклонения V, а только одного из локальных минимумов.

2. Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.

3. Число кластеров надо знать заранее.

K-means с разными метриками



Взаимосвязь кластерного и регрессионного анализа

Кластеризация переменных

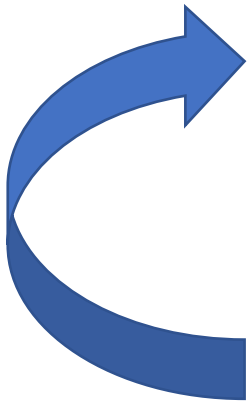
Метрика расстояния: коэффициент корреляции

Цель:

1. Идентификация характерных переменных или переменных, которые вносят уникальный вклад в данные;
2. уменьшение числа переменны (замена переменных на кластерные компоненты).

Этапы кластерного анализа

1. Отбор выборки для кластеризации
2. Определение множества переменных, по которым будут оцениваться объекты в выборке
3. Удаление выбросов, борьба с пропущенными значениями
4. Стандартизация данных
5. Вычисление значений той или иной меры сходства между объектами.
6. Применение метода кластерного анализа для создания групп сходных объектов
7. Интерпретация результатов
8. Проверка достоверности результатов кластерного решения



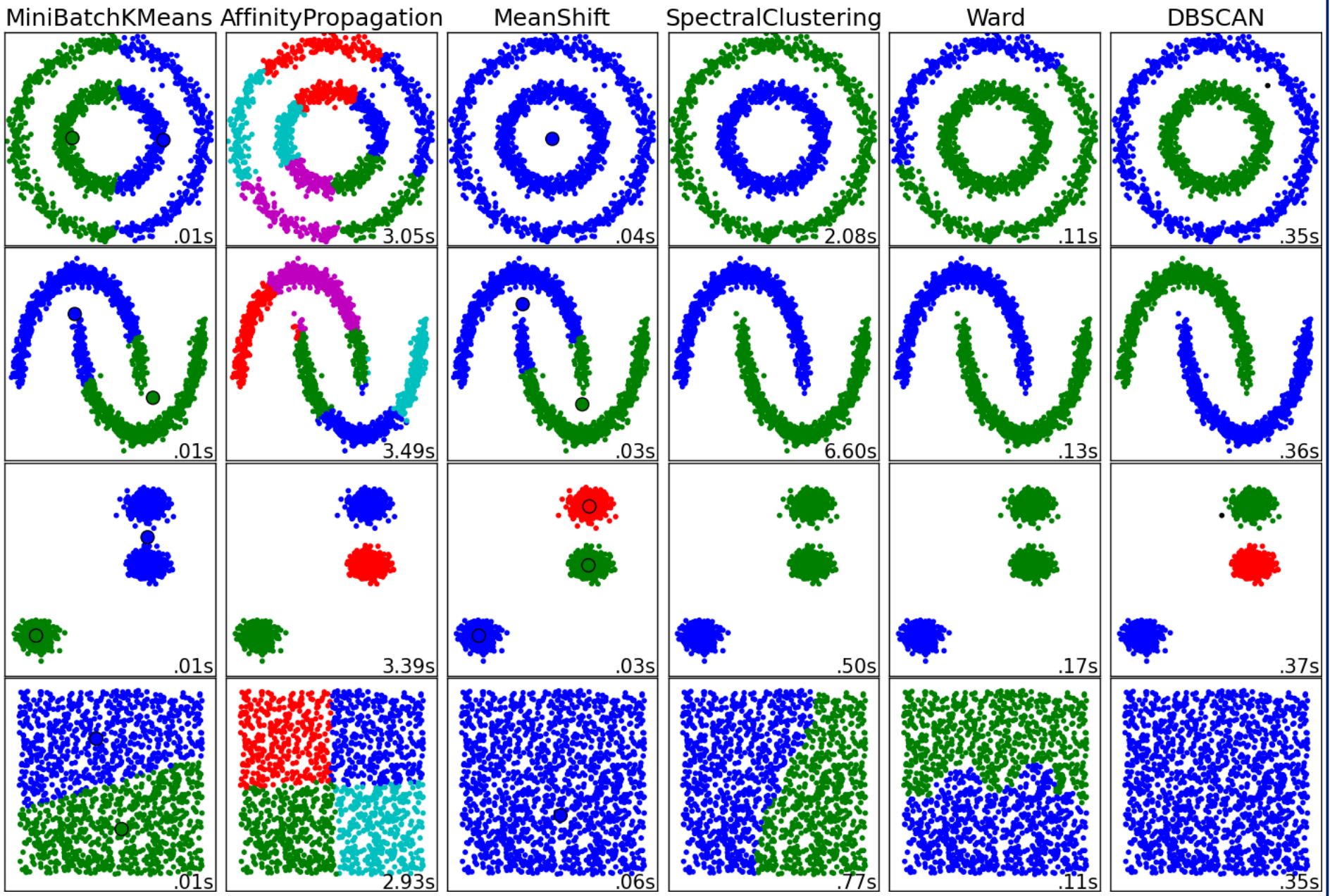
Основные требования к данным при кластерном анализе

1. Переменные не должны коррелировать между собой переменные должны быть безразмерными

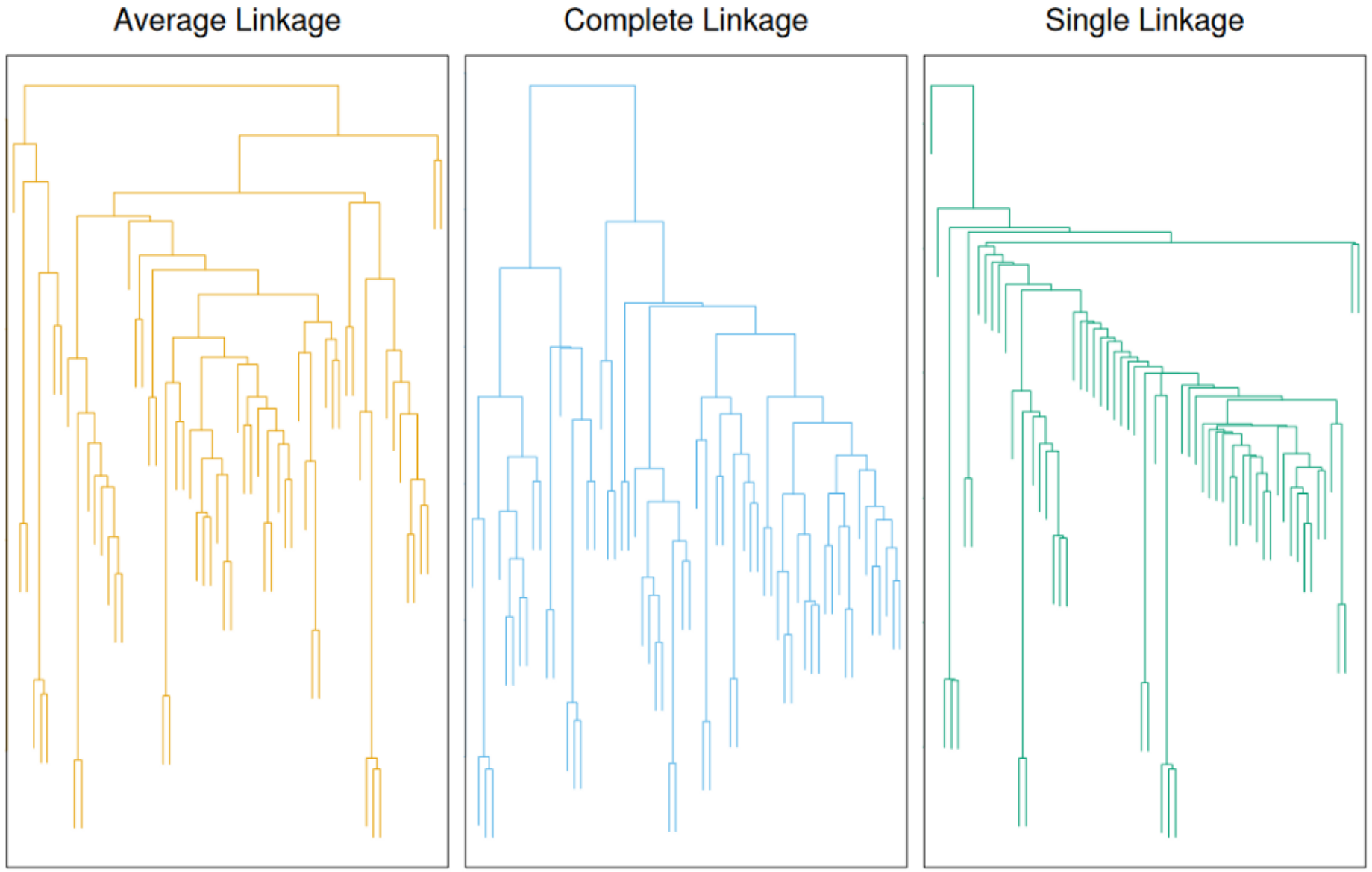
2. Распределение переменных должно быть близко к нормальному показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов

3. Выборка должна быть однородна, не содержать «выбросов»

Кластеризация разными методами и метриками – разные результаты!



Кластеризация разными методами и метриками – разные результаты!



Особенности кластерного анализа

- Кластеризация не приносит каких-либо результатов без содержательной интерпретации каждого кластера.
- Интерпретация предполагает присвоение каждому кластеру емкого названия, отражающего его суть, например «Пожилые женщины с факторами риска», «Пациенты, ведущие неактивный образ жизни» и т. д.
- Для интерпретации аналитик детально исследует каждый кластер: **его статистические характеристики, распределение значений признаков объекта в кластере, оценивает мощность кластера — число объектов, попавших в него**.
- Интерпретация значительно облегчается, если имеются способы представления результатов кластеризации в специализированном виде: ***дендограммы, кластерограммы, карты***.

Общие выводы

1. Не существует однозначно наилучшего критерия качества кластеризации.

2. Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием.

3. Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

Кластерный анализ на Python

Библиотека Scikit-Learn

Method name
K-Means
Affinity propagation
Mean-shift
Spectral clustering
Ward hierarchical clustering
Agglomerative clustering
DBSCAN
OPTICS
Gaussian mixtures
BIRCH

Название метода	Параметры	Масштабируемость	Использование	Геометрия (используемая метрика)
K-средник	число кластеров	Очень большое значение **n симплов** среднее **n_clusters** вместе с Мини батчи K-средних	Универсальный, любой размер кластеров, плоская геометрия, не слишком много кластеров	Дистанция между точками
Афинное распространение	дамфинг, предпочтение выборки	Не масштабируется с помощью **n_clusters**	Много кластеров, не равномерный размер кластера, неплоская геометрия	Дистанция графа (например граф ближайших соседей)
Средний сдвиг	пропускная способность	Не масштабируется с помощью **n_clusters**	Много кластеров, не равномерный размер кластера, неплоская геометрия	Дистанция между точками
Спектральная кластеризация	число кластеров	Средняя n симплов маленькое **n_clasters**	Мало кластеров, или размер кластера, неплоская геометрия	Дистанция графа (например граф ближайших соседей)
Иерархическая кластеризация	чилсо кластеров или порог расстояния	Большое n симплов и **n_clasters**	Мало кластеров, возможно ограничене связей	Дистанция между точками
Агломеративная кластеризация	число кластеров, порог дистанции, тип связи, дистанция	Большое n симплов и **n_clasters**	Мало кластеров, возможно ограничене связей и не Евклидовое расстояние	Любая попарная дистанция
DBSCAN	размер окрестности	Очень большое n симплов и среднее **n_clasters**	не плоская геометрия, неравномерный размер кластеров	Дистанция между ближайшими точками
OPTICS	Минимальное количество элементов в кластере	Очень большое n симплов и большое **n_clasters**	не плоская геометрия, неравномерный размер кластеров, переменная плотность кластеров	Дистанция между точками
Гауссовская Смешаянная модель	много параметров	не масштабируемо	плоская геометрия, подходит для оценки плотности	Расстояния до центров Махаланобиса
Birch	факторы ветвления, порог, не обязательный глобальный	Большое n симплов и **n_clasters**	Большой объем данных, удаление выбросов, сокращение данных	Евклидовое расстояние между точками

https://scikit-learn.org/stable/modules/clustering.html

Набор Ирисы Фишера библиотеки scikit-learn

Один из «традиционных» наборов данных (dataset), используемых для обучения и содержащий данные по трём видам цветов ириса: Ирис щетинистый (Iris setosa), Ирис виргинский (Iris virginica) и Ирис разноцветный (Iris versicolor) В библиотеке scikit-learn есть и другие датасеты:

```
from sklearn import datasets
df = datasets.load_iris()
```

Ирисы Фишера				
Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка	Вид ириса
5.1	3.5	1.4	0.2	*setosa*
4.9	3.0	1.4	0.2	*setosa*
4.7	3.2	1.3	0.2	*setosa*
4.6	3.1	1.5	0.2	*setosa*
5.0	3.6	1.4	0.2	*setosa*
5.4	3.9	1.7	0.4	*setosa*
4.6	3.4	1.4	0.3	*setosa*
5.0	3.4	1.5	0.2	*setosa*
4.4	2.9	1.4	0.2	*setosa*
4.9	3.1	1.5	0.1	*setosa*
5.4	3.7	1.5	0.2	*setosa*
4.8	3.4	1.6	0.2	*setosa*
4.8	3.0	1.4	0.1	*setosa*

Iris setosa

Iris virginica

Iris versicolor

Задание до следующего занятия (29.04.2024)

1. Доделать практическое задание
2. Подготовиться к тестированию по данной лекции