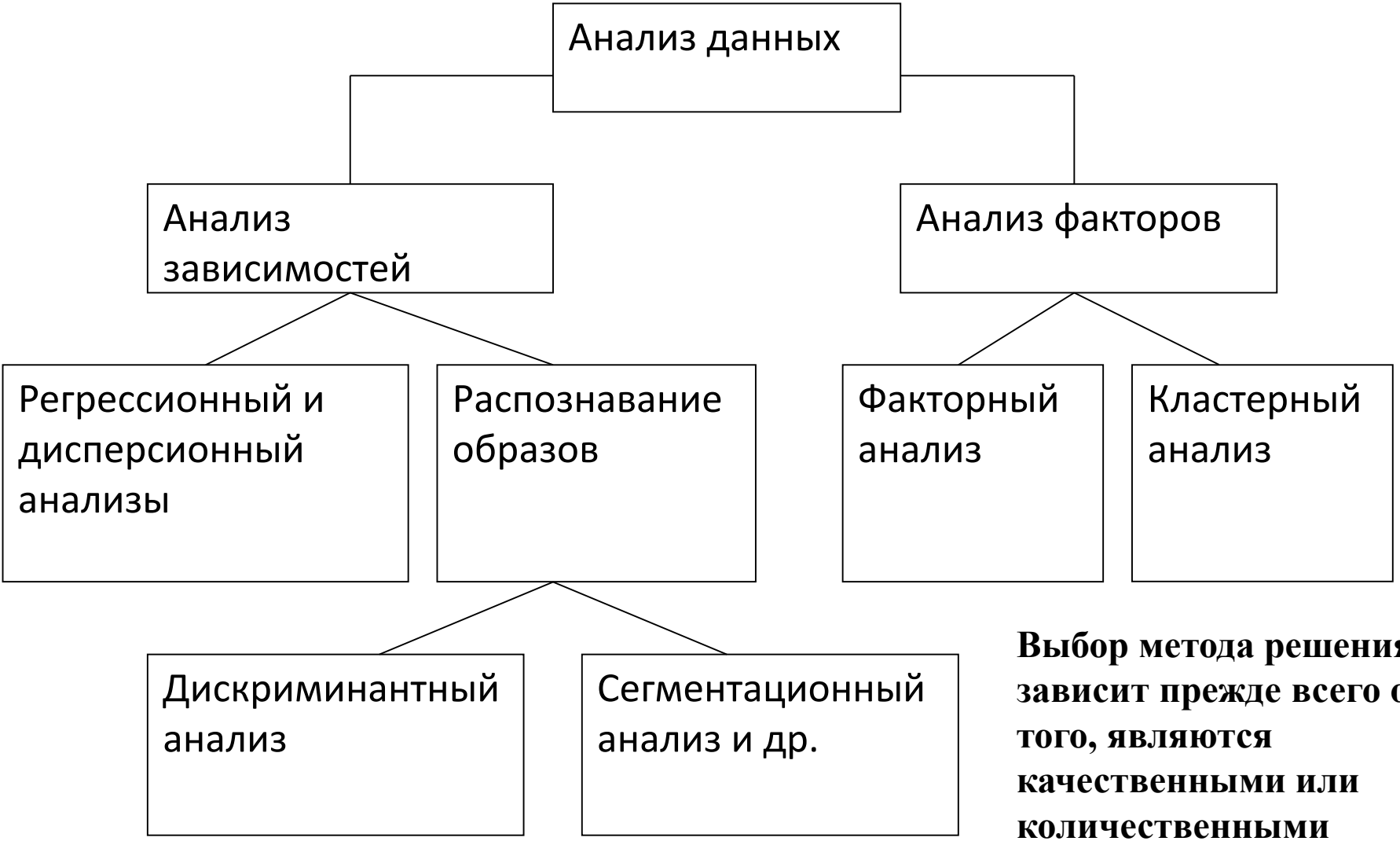
Лекция 4-5

МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА

Классификация методов анализа данных



Выбор метода решения зависит прежде всего от того, являются качественными или количественными зависимые переменные.

Окончательно решение о выборе метода анализа данных принимается в зависимости от типа независимых переменных.

Кластерный анализ. Основные понятия

Для поиска качественных факторов применяется группа методов, известная под названием ***кластерный анализ***.

Методы кластерного анализа позволяют разбить изучаемую совокупность объектов на группы «схожих» объектов, называемых кластерами.

Методы кластеризации довольно разнообразны, в них по-разному выбирается способ определения близости между кластерами (и между объектами), а также используются различные алгоритмы вычислений.

Поэтому результаты кластеризации зависят от выбранного метода.

Особенности кластерного анализа

В отличие от задач классификации, ***кластерный анализ*** не требует априорных предположений о наборе данных, не накладывает ограничения на *представление* исследуемых объектов, позволяет анализировать показатели различных типов данных (*интервальным данным, частотам, бинарным данным*).

При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

Кластерный анализ позволяет сокращать *размерность* данных, делать ее наглядной.

Особенности кластерного анализа

Кластерный анализ может применяться к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой.

Кластерный анализ параллельно развивался в нескольких направлениях, таких как биология, психология, др., поэтому у большинства методов существует по два и более названий. Это существенно затрудняет работу при использовании кластерного анализа.

История возникновения

1. Концепция классификации и систематизации, предложенная французским ботаником **Огюстеном Декандолем (1778-1841) в 1813** году с целью систематизации растений. Данная теория получила наименование таксономия.

2. Статья польского антрополога **Яна Чекановского**, которую он написал **в 1911 году**. В своей работе он показывает идею «структурной классификации», содержащую главную мысль кластерного анализа – выделение компактных групп близких объектов, а так же некоторые методы выделения таких групп, которые лежат в основе более последних алгоритмов.

3. «Метод корреляционных плеяд», созданный советским гидробиологом **П.В. Терентьевым в 1925 году**. Однако издан он был лишь через много лет в **1959 г.**

4. Сам **термин «кластерный анализ»** был впервые введен и использован только **в 1939 году английским ученым Р. Трионом**

Кластерный анализ. Основные задачи

Цели кластеризации могут быть различными в зависимости от особенностей конкретной прикладной задачи.

- **Упростить дальнейшую обработку данных**, разбить множество на группы схожих объектов чтобы работать с каждой группой в отдельности.
- **Сократить объём хранимых данных**, оставив по одному представителю от каждого кластера (задачи сжатия данных).
- **Выделить нетипичные объекты**, которые не подходят ни к одному из кластеров.
- **Построить иерархию множества объектов**.

Кластерный анализ. Основные понятия

Задачи кластерного анализа можно объединить в следующие группы:

1. Разработка типологии или классификации.

2. Исследование полезных концептуальных схем группирования объектов.

3. Представление гипотез на основе исследования данных.

4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Кластерный анализ. Основные понятия

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

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*.

Размер кластера может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера.

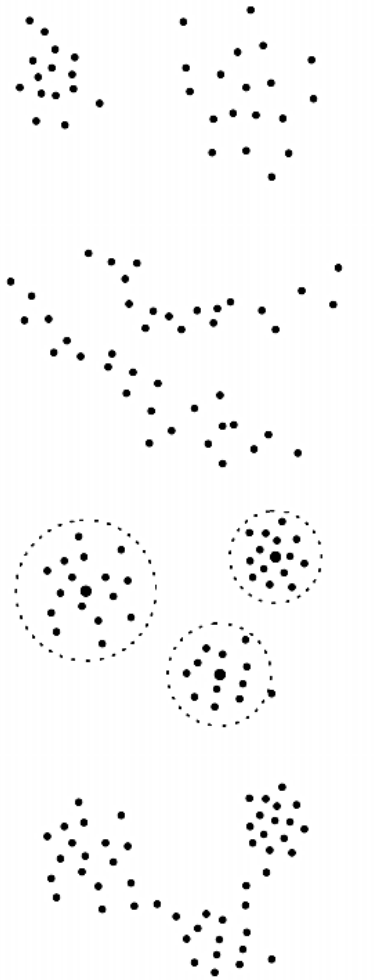
Кластерный анализ. Основные понятия

Для определения сходства («близости») объектов используются различные метрики.

Спорный объект - это *объект*, который по мере сходства может быть отнесен к нескольким кластерам.

Объект относится к кластеру, если *расстояние* от объекта до *центра кластера* меньше *радиуса кластера*. Если условие выполняется для двух и более кластеров, объект является спор ным.

Типы кластерных структур*



Сгущения: внутрикластерные расстояния, как правило, меньше межкластерных.

Ленты: для любого объекта найдётся близкий к нему объект того же кластера, в то же время существуют объекты одного кластера, которые не являются близкими.

Кластеры с центром: в каждом кластере найдётся объект, такой, что почти все объекты кластера лежат внутри шара с центром в этом объекте.

Кластеры могут соединяться перемычками, что затрудняет работу многих алгоритмов кластеризации.

* Источник: Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: http://www.ccas.ru/voron

Типы кластерных структур*

Кластеры могут накладываться на разреженный фон из редких нетипичных объектов.

Кластеры могут перекрываться.

Кластеры могут образовываться не по принципу сходства, а по каким-либо иным, заранее неизвестным, свойствам объектов. Стандартные методы кластеризации здесь бессильны.

Кластеры могут вообще отсутствовать. В этом случае надо применять не кластеризацию, а иные методы анализа данных.

*Источник: Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: http://www.ccas.ru/voron

Кластерный анализ. Основные понятия

Понятие «расстояние между объектами» является интегральной мерой сходства объектов между собой.

Расстоянием между объектами в пространстве признаков называется такая величина d_{ij} , которая удовлетворяет следующим аксиомам:

$d_{ij} > 0$ (неотрицательность расстояния)

$d_{ij} = d_{ji}$ (симметрия)

$d_{ij} + d_{jk} > d_{ik}$ (неравенство треугольника)

Если d_{ij} не равно 0, то i не равно j (различимость нетождественных объектов)

Если $d_{ij} = 0$, то $i = j$ (неразличимость тождественных объектов)

Меру близости (сходства) объектов удобно представить как обратную величину от расстояния между объектами.

Кроме термина **"расстояние"** в литературе часто встречается и другой термин - "метрика", который подразумевает метод вычисления того или иного конкретного расстояния.

Метрики кластеризации

1) **евклидово расстояние или евклидова метрика** $d_{ij} = \left(\sum_{k=1}^{v} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$

2) квадрат евклидова расстояния $d^2_{ij} = \sum_{k=1}^{v} (x_{ik} - x_{jk})^2$

3) **степенное расстояние**

(обобщенная метрика Минковского) $d_{ij} = (\sum_{k=1}^{v} |x_{ik} - x_{jk}|^p)^{\frac{1}{p}}$

4) расстояние городских кварталов

(манхэттенское расстояние, city-block) $d_{ij} = \sum_{k=1}^{v} |x_{ik} - x_{jk}|$

5) метрика доминирования

(Sup-метрика, $d_{ij} = (\sum_{k=1}^{v} |x_{ik} - x_{jk}|^{\infty})^{\frac{1}{\infty}}$

расстояние Чебышева $d_{ij} = max| x_{ik} - x_{jk}|.$

Метрики кластеризации

6) **расстояние Махаланобиса** $d_{ij} = (X_i - X_j)^T S^{-1} (X_i - X_j)$

В отличие от метрики Минковского и евклидовой метрики, расстояние Махаланобиса через матрицу дисперсий-ковариаций S связано с корреляциями переменных.

Когда корреляции между переменными равны нулю, расстояние Махаланобиса эквивалентно квадрату евклидового расстояния.

7) **расстояние Хемминга** $d_{ij} = \sum_{k=1}^{v} |x_{ik} - x_{jk}|$

Применяется в случае использования дихотомических (имеющих всего два значения) качественных признаков, равно числу несовпадений значений соответствующих признаков для рассматриваемых i-того и j-того объектов.

От выбора метрики во многом зависит результат кластеризации.
Выбор осуществляется в зависимости от пространства, в котором расположены объекты и неявных характеристик кластеров.

Общая методология кластеризации

[Jain, Dubes. Algorithms for clustering data (Prentice-Hall, 1988)]

Общая методология кластеризации

Сбор данных: получение «сырых» данных из различных источников.

Первоначальный отбор: подготовка данных к анализу, нормализация. Выявление данных, которые будут мешать дальнейшему анализу, например, незначащие характеристики, дубликаты, противоречия.

Представление: перевод данных в форму, пригодную для дальнейшего анализа.

Тенденция кластеризации: выявление неслучайной структуры в данных. Если данные не имеют тенденцию к кластеризации, то выбирается другая техника анализа данных.

Стратегия кластеризации: выбор соответствующего метода (иерархический\неиерархический) и затем алгоритма. Внимание должно быть уделено соответствию алгоритма конкретным данным.

Валидация: сравнение с данными, полученными «извне»; сравнение с данными, полученными при работе других алгоритмов.

Интерпретация: графическое представление результатов кластерного анализа.

Этапы кластеризации

- Выявление вектора характеристик: выбор наиболее эффективных подмножеств характеристик или создание новых характеристик путем трансформации существующих.
- Выбор метрики: выбор меры расстояний для определения «близости» объектов. Выбор осуществляется в зависимости от пространства, в котором расположены объекты и неявных характеристик кластеров.
- Разбиение объектов на кластеры: выполняется в соответствии с выбранным алгоритмом. Производится изменение метрики или вектора характеристик при неудовлетворительном результате разбиения.

Кластерный анализ. Классификация методов

Кластерный анализ. Классификация методов

Иерархические (Hierarchical) — построение дендограммы (дерево вложенных кластеров)

Агломеративные (Aglomerative) — в начале работы алгоритма количество кластеров равно количеству объектов, далее итерационно «снизу-вверх» ближайшие два объединяются

Дивизимные (Divisive) — в начале работы алгоритма все объекты относятся к одному кластеру, далее итерационно «сверху-вниз» каждый кластер делится на два

Дендрограмма (dendrogram)

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.

Дендрограмма (*dendrogram*) - древовидная диаграмма, содержащая ***n*** уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

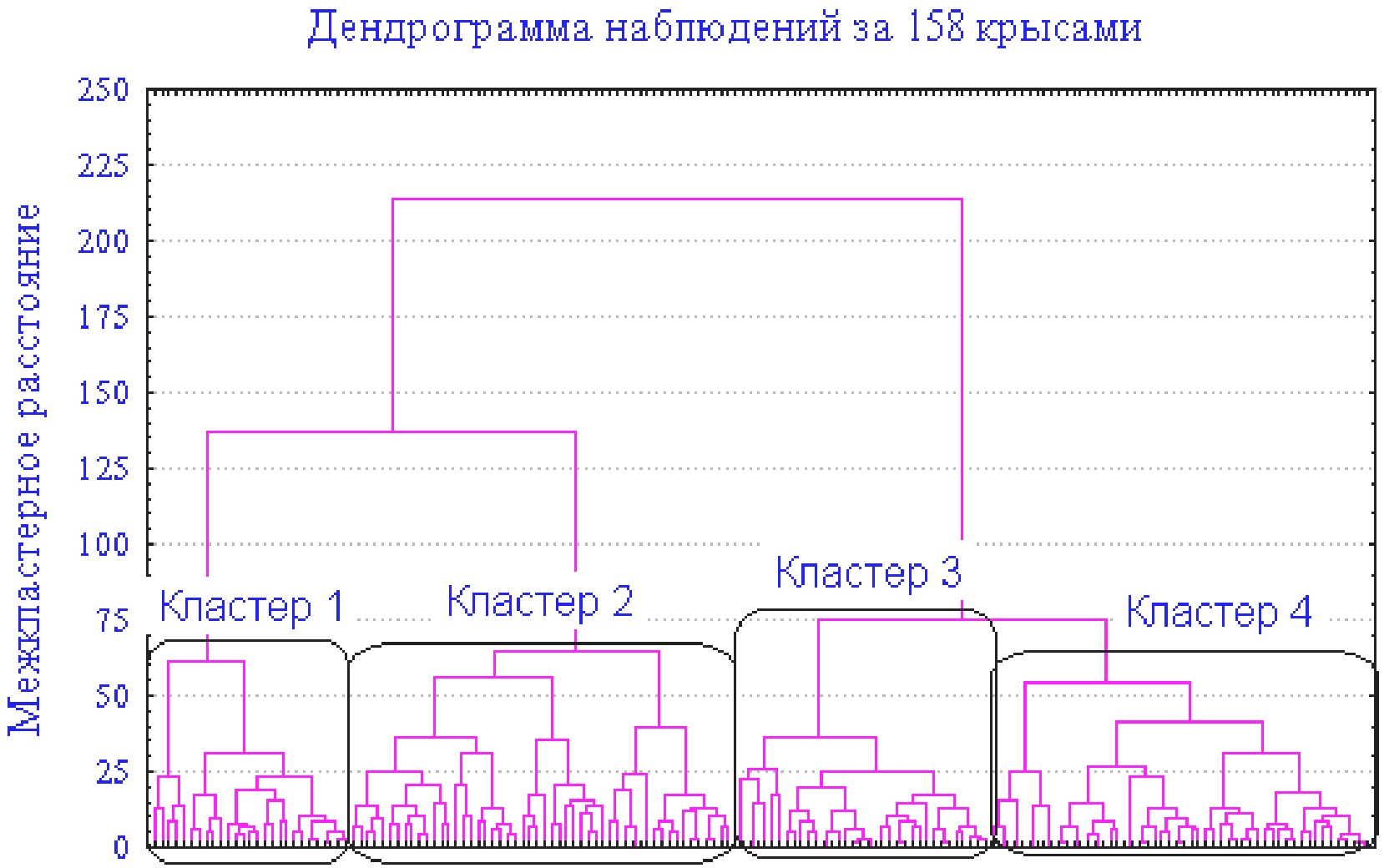
Дендрограмма (dendrogram)

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

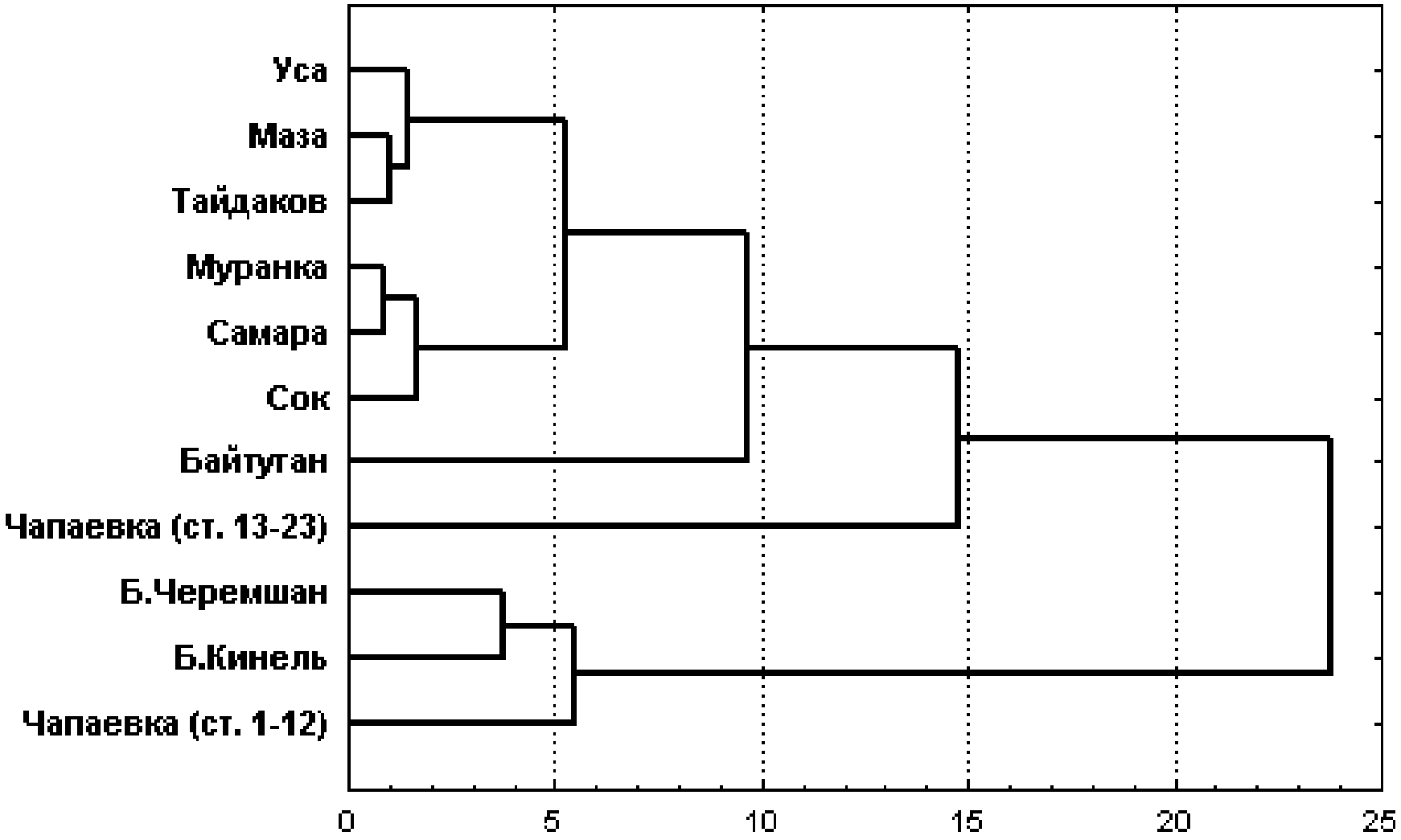
Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

Существует много способов построения *дендрограмм*. В *дендрограмме* объекты могут располагаться вертикально или горизонтально.

Пример горизонтальной дендрограммы



Пример вертикальной дендрограммы (Кластеризация рек Самарской области)



Методы связи:

включают методы «ближайшего соседа», «далекого соседа» и «среднего расстояния»

Метод «ближайшего соседа» (одиночная связь) первыми объединяют два объекта, расстояние между которыми минимально.

Далее определяют следующее по величине самое короткое расстояние, и в кластер с первыми двумя объектами вводят третий объект.

Расстояние между кластерами – расстояние между их ближайшими точками.

В *методе «далекого соседа» (полная связь)* расстояния между кластерами вычисляют как расстояния между их самыми удаленными точками

В *методе «среднего расстояния»* расстояние между кластерами определяют, как среднее значение всех расстояний между объектами двух кластеров.

Метод Ворда

В этом методе ***в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений***, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект.

На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов.

Этот метод направлен на объединение близко расположенных кластеров.

MST (Minimum Spanning Trees)

- ***иерархический дивизимный*** алгоритм, основанный на теории графов.

Основная идея алгоритма: строится минимальное остовное дерево на основе множества всех исходных объектов, далее дерево делится на кластеры.

> Минимальное остовное дерево (или минимальное покрывающее дерево) в связанном, взвешенном, неориентированном графе — это остовное дерево этого графа, имеющее минимальный возможный вес, где под весом дерева понимается сумма весов входящих в него рёбер.

Достоинства: алгоритм выделяет кластеры произвольной формы.

Алгоритм:

1.Построение минимального остовного дерева (алгоритмы Борувки, Крускала, Прима).

2.Разделение на кластеры. Дуги с наибольшими весами разделяют кластеры.

Кластерный анализ. Классификация методов

- Неиерархические (Partitional)
- Вероятностные
- k-средних (k-means)
- k-медиан и k-метоидов (k-medians, k-medoids)
- Плотностные (метод поиска сгущений)

Алгоритм K-means (*k* внутригрупповых средних) Мак-Куина (McQueen)

Простой и широко используемый *неиерархический* алгоритм кластеризации.

Основная идея алгоритма: разбиение всего множества объектов на *k* кластеров с последующим пересчетом их центров и перераспределением объектов по кластерам.

Значение k - количество кластеров задается заранее и является основным из входных данных алгоритма.

Достоинства: простота использования; быстрота использования; понятность и прозрачность алгоритма.

Недостатки: качество результата сильно зависит от выбора начального разбиения; медленная работа на больших объемах исходных данных; необходимо задавать количество кластеров; алгоритм чувствителен к выбросам.

Алгоритм k-средних

1. Выбирается k объектов (по числу кластеров) и на первом шаге они считаются в качестве центров.

2. В соответствии с выбранной метрикой каждый объект исходного множества присваивается определенному кластеру, исходя из близости его к центру кластера.

3. Пересчитываются центры кластеров в соответствии с влиянием новых объектов, попавших в кластер.

4. Далее Пункт 2.

5. Алгоритм заканчивается когда кластерные центры стабилизировались или число итераций стало равно максимальному числу итераций.

Каждый объект x из $X = R_n$ описывается n числовыми признаками: $x \equiv f_1(x), \ldots, f_n(x)$. Каждый кластер $y \in Y$ описывается n-мерной гауссовской плотностью $p_y(x)$ с центром $\mu_y = (\mu_{y1}, \ldots, \mu_{yn})$ и диагональной матрицей ковариаций Σ_y.

ИЛИ:

1) сформировать начальное приближение центров всех кластеров $y \in Y$:
μ_y – наиболее удалённые друг от друга объекты выборки;

2) повторять

- отнести каждый объект к ближайшему центру:

$$y_i := \arg\min_{y \in Y} \rho(x_i; \mu_y), i = 1 \ldots l;$$

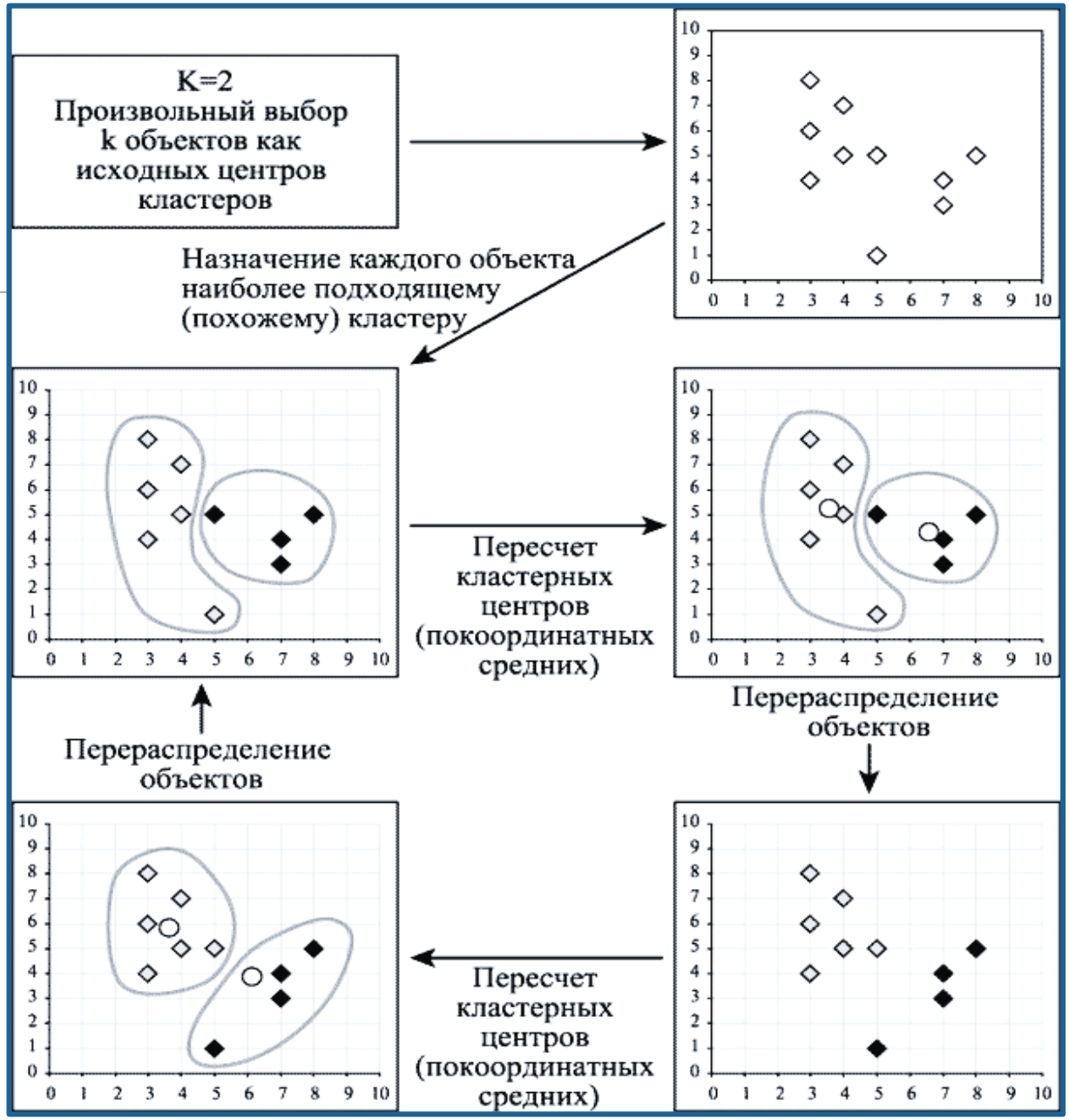
- вычислить новое положение центров:

$$\mu_{yj} := \frac{\sum_{i=1}^{l} [y_i = y] f_j(x_i)}{\sum_{i=1}^{l} [y_i = y]}, y \in Y, j = 1, \ldots, n;$$

пока y_i не перестанут изменяться *.

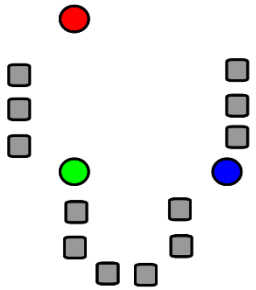
* Источник: Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: http://www.ccas.ru/voron

Пример работы алгоритма k-средних (k=2)

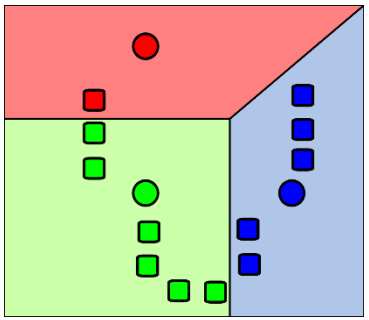


Пример работы алгоритма k-средних (k=3)

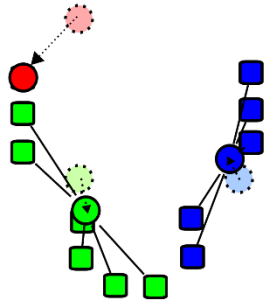
1. Исходные точки и случайно выбранные начальные точки.



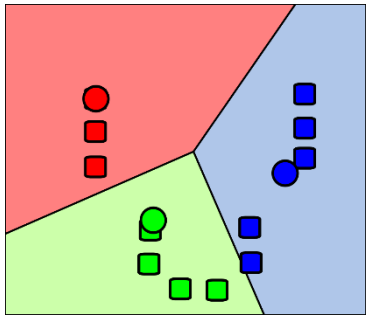
2. Точки, отнесённые к начальным центрам.



3. Вычисление новых центров кластеров



4. Предыдущие шаги, за исключением первого, повторяются, пока алгоритм не сойдётся.



Разбиение на плоскости – диаграмма Вороного относительно начальных центров.

Другие неиерархические методы

Метод k-медиан - вариация метода k-средних для задач кластеризации, где для определения центроида кластера вместо среднего вычисляется *медиана*.

- Задача определения k-медиан состоит в поиске таких k-центров, что сформированные по ним кластеры будут наиболее «компактными».
- Формально, при заданных точках данных x_i, k центров должны быть выбраны так, чтобы минимизировать сумму расстояний от каждой x_i до ближайшего j-го центроида.
- Метод иногда работает лучше, чем метод k-средних, где минимизируется сумма квадратов расстояний.

Другая альтернатива - **метод k-медоидов**, в котором ищут оптимальный *медоид*, а не медиану кластера (медоид является одной из точек данных, в то время как медианы таковыми быть не обязаны).

Другие неиерархические методы

Метод поиска сгущений. Суть итеративного алгоритма данного метода – в применении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков с целью поиска локальных сгущений объектов.

Метод поиска сгущений требует вычисления *матрицы расстояний* (или *матрицы мер сходства*) между объектами и выбора первоначального центра сферы.

- Обычно на первом шаге центром сферы служит объект (точка), в ближайшей окрестности которого расположено наибольшее число соседей.
- На основе заданного радиуса сферы (R) определяется совокупность точек, попавших внутрь этой сферы, и для них вычисляются координаты центра (вектор средних значений признаков).

Другие неиерархические методы

- Когда очередной пересчет координат центра сферы приводит к такому же результату, как и на предыдущем шаге, перемещение сферы прекращается, а точки, попавшие в нее, образуют кластер, и из дальнейшего процесса кластеризации исключаются.
- Перечисленные процедуры повторяются для всех оставшихся точек.
- Работа алгоритма завершается за конечное число шагов, и все точки оказываются распределенными по кластерам.
- Число образовавшихся кластеров заранее неизвестно и сильно зависит от радиуса сферы.
- Для оценки устойчивости полученного разбиения целесообразно повторить процесс кластеризации несколько раз для различных значений радиуса сферы, изменяя каждый раз радиус на небольшую величину.

Качество кластеризации

— степень приближения результата кластеризации к идеальному решению. Для большинства задач идеальное решение неизвестно.

Оценка качества кластеризации может быть произведена двумя способами:

- ***Формальный способ***. Формальный способ основан на определении формальных критериев. Наилучшим считается решение, для которого значение формального критерия максимально.
- ***Экспертный способ***. Решение оценивается специалистами заданной предметной области.

Качество кластеризации

Критерии качества:

- **Показатели четкости**: коэффициент разбиения, модифицированный коэффициент разбиения, индекс четкости.
- **Энтропийные критерии**: энтропия разбиения, модифицированная энтропия.
- **Показатель компактности и изолированности**
- **Индекс эффективности**

Основные этапы оценки качества кластеризации:

1.Алгоритм кластеризации, построение модели данных.

2.Вычисление критерия качества кластеризации. Критерии вычисляются на основе получившейся в ходе работы алгоритма кластеризации матрицы принадлежности и/или множества кластерных центров.

3.Определение параметров настройки алгоритма.

Кластерный анализ. Сравнение методов

Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации.

- *Недостаток* - аналитик должен заранее определить количество кластеров, количество итераций или правило остановки, а также некоторые другие параметры кластеризации.

Если нет предположений относительно числа кластеров, рекомендуют использовать **иерархические алгоритмы**.

Кластерный анализ. Сравнение методов

Однако если объем выборки не позволяет это сделать, возможный путь - проведение ряда экспериментов с различным количеством кластеров, например, начать разбиение совокупности данных с двух групп и, постепенно увеличивая их количество, сравнивать результаты.

За счет такого "варьирования" результатов достигается достаточно большая гибкость кластеризации.

Иерархические методы, в отличие от неиерархических, отказываются от определения числа кластеров, а строят полное дерево вложенных кластеров.

Сложности иерархических методов кластеризации: ограничение объема набора данных; выбор меры близости; негибкость полученных классификаций.

Кластерный анализ. Сравнение методов

- Преимущество этой группы методов в сравнении с неиерархическими методами - их наглядность и возможность получить детальное представление о структуре данных.

- При использовании ***иерархических методов*** существует возможность достаточно легко идентифицировать выбросы в наборе данных и, в результате, повысить качество данных.

- Эта процедура лежит в основе **двухшагового алгоритма** кластеризации. Такой набор данных в дальнейшем может быть использован для проведения ***неиерархической кластеризации.***