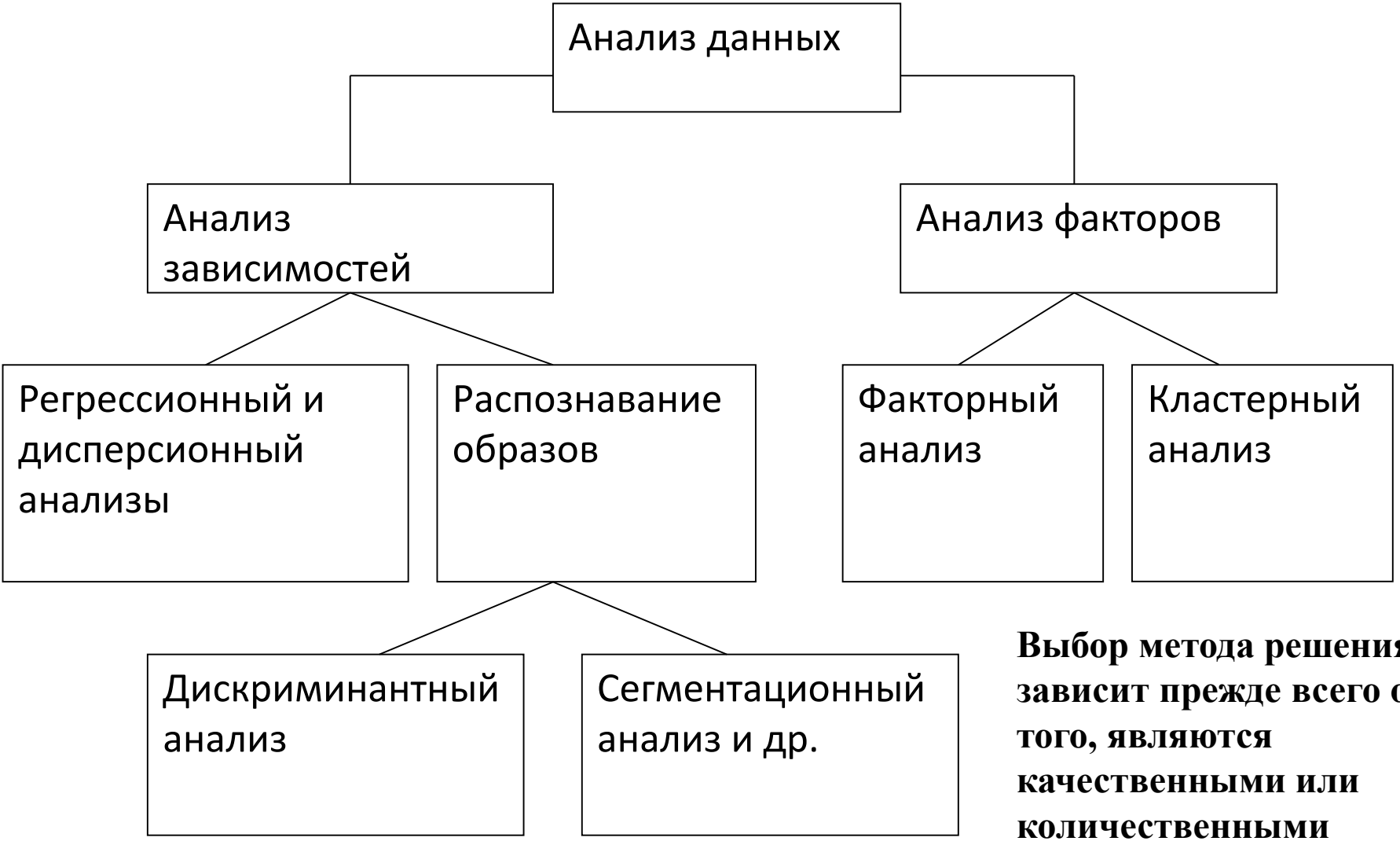
Лекция 4

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.

Синонимами термина **«кластеризация»** являются «автоматическая классификация», «неконтролируемая классификация», «обучение без учителя» и «таксономия» (taxonomy).

Особенности кластерного анализа

Алгоритмы кластеризации являются в большой степени *эвристическими*.

Эвристический алгоритм – это алгоритм решения задачи, правильность которого для всех возможных случаев не доказана, но про который известно, что он даёт достаточно хорошее решение в большинстве случаев.

В действительности может быть даже известно, что эвристический алгоритм формально неверен.

Его всё равно можно применять, если при этом он даёт неверный результат только в отдельных, достаточно редких и хорошо выделяемых случаях или же даёт неточный, но всё же приемлемый результат.

Проще говоря, эвристика – это не полностью математически обоснованный (или даже «не совсем корректный»), но при этом практически полезный алгоритм.

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*.

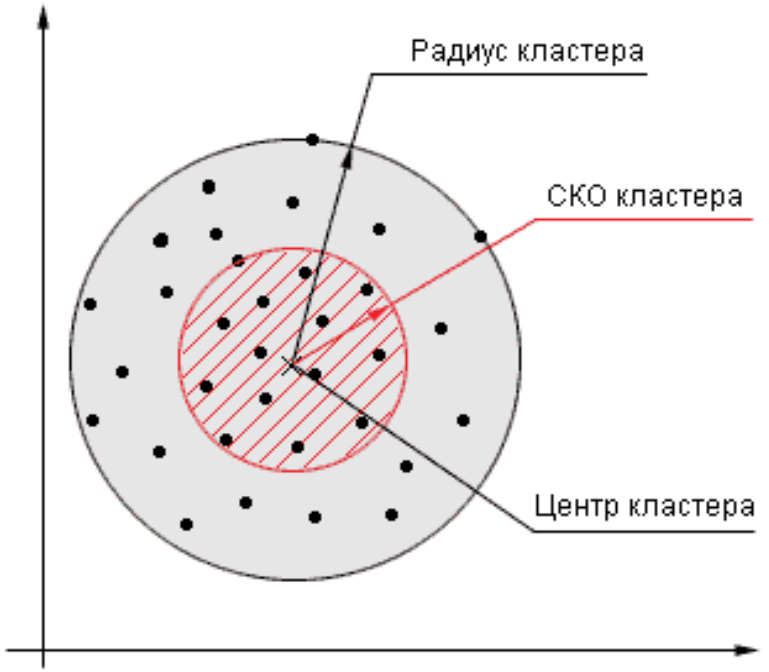
Размер кластера может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера.

Кластерный анализ. Основные понятия

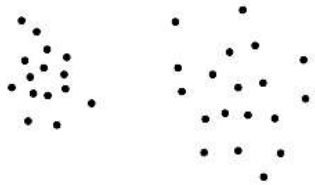
Для определения сходства («близости») объектов используются различные метрики.

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

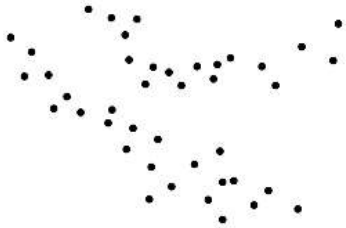
Объект относится к кластеру, если *расстояние* от объекта до *центра кластера* меньше *радиуса кластера*. Если условие выполняется для двух и более кластеров, объект является спор ным.



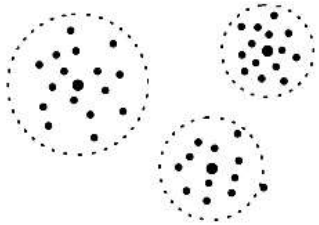
Типы кластерных структур*



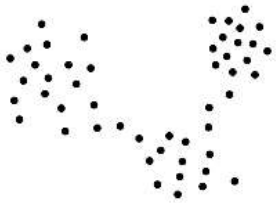
Сгущения: внутрикластерные расстояния, как правило, меньше межкластерных.



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



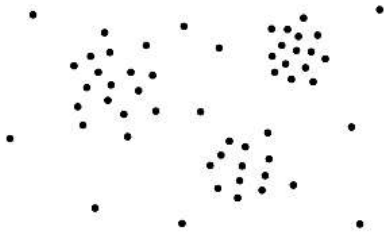
Кластеры с центром: в каждом кластере найдётся объект, такой, что почти все объекты кластера лежат внутри шара с центром в этом объекте.



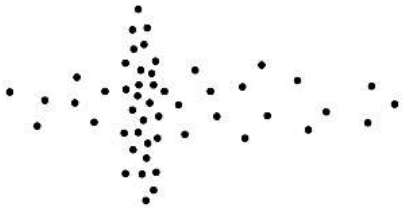
Кластеры могут соединяться перемычками, что затрудняет работу многих алгоритмов кластеризации.

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

Типы кластерных структур*

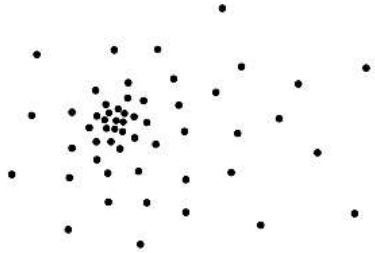


Кластеры могут накладываться на разреженный фон из редких нетипичных объектов.



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.