



# Технология Data Mining

Лесничин Антон,  
ПИЭМ-191

# Краткая характеристика основных методов Data Mining

Классификация по принципу работы с данными разбивает методы Data Mining на две крупные категории:

- методы, связанные с непосредственным использованием (сохранением) данных. Данные в ходе обработки детализируются при построении прогностической модели или в ходе анализа исключений. Однако такие методы малоэффективны при работе с крупными массивами данных. Методики этой категории применяются в формах кластерного анализа, метода ближайшего соседа, метода k-ближайшего соседа, рассуждений по аналогии.
- дистилляция шаблонов - формирование и применение закономерностей, имеющих упорядоченный вид, то есть извлечение информации из изначальных данных с ее преобразованием в определенную систематизированную конструкцию.

Технологии этой группы представлены логическими, визуализирующими, кросс-табуляционными и базирующимися на уравнениях методами. Задействование этих методов обеспечивает эффективное применение полученных в ходе свободного поиска результатов (они более компактны по сравнению с базами данных) и преобразование этих сведений в понятные для пользователей закономерности.

# Краткая характеристика основных методов Data Mining

В свою очередь, способы логической аналитики делятся на подклассы, к которым относятся постановка нечетких запросов, использование символьных правил, деревьев решений и генетических алгоритмов. Технологии кросс-табуляции основаны на применении так называемых агентов, байесовских сетей и визуальных кросс-таблиц. Статметоды и нейронные сети объединяются в методы на основе уравнений.

Существует еще одна разбивка методов Data Mining - по принципам применения математических моделей в обучении. Здесь выделяются две группы:

- статистические методы, в которых используется усредненный опыт по данным, накопившимся в БД за длительный период. При использовании статметодов предварительно анализируется природа статистических данных, выявляются связи и закономерности, осуществляется многомерный статистический анализ, строятся динамические модели и прогноз на основе временных рядов;
- кибернетические методы, в которых используются основы компьютерной математики и технологии искусственного интеллекта. В число таких методов входят: эволюционное программирование, нейросети, системы обработки экспертных знаний.

\*К кибернетическим методам также относятся ассоциативные правила, деревья решений, нечеткая логика, генетические алгоритмы.

# Методы классификации и кластеризации

Оценивание классификационных методов

Оценивание методов следует проводить, исходя из следующих характеристик: скорость, робастность, интерпретируемость, надежность.

- Скорость характеризует время, которое требуется на создание модели и ее использование.
- Робастность, т.е. устойчивость к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями в данных.

Интерпретируемость обеспечивает возможность понимания модели аналитиком.

- Свойства классификационных правил:
- размер дерева решений;
- компактность классификационных правил.

Надежность методов классификации предусматривает возможность работы этих методов при наличии в наборе данных шумов и выбросов.

# Методы классификации и

## кластеризации

### Задача кластеризации

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не predetermined.

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек".

**Цель кластеризации** - поиск существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Характеристиками кластера можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

Вопрос, задаваемый аналитиками при решении многих задач, состоит в том, как организовать данные в наглядные структуры, т.е. развернуть таксономии.

Наибольшее применение кластеризация первоначально получила в таких науках как биология, антропология, психология. Для решения экономических задач кластеризация длительное время мало использовалась из-за специфики экономических данных и явлений.

# Сравнительная таблица классификации и кластеризации

## Характеристика

Контролируемость обучения  
Стратегия  
Наличие метки класса

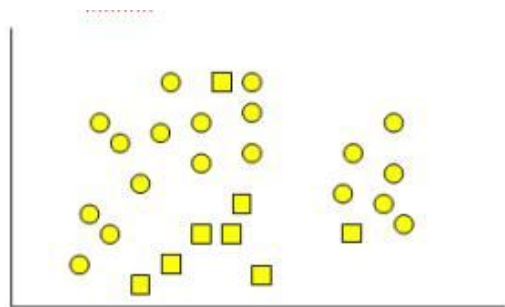
## Основание для классификации

## Классификация

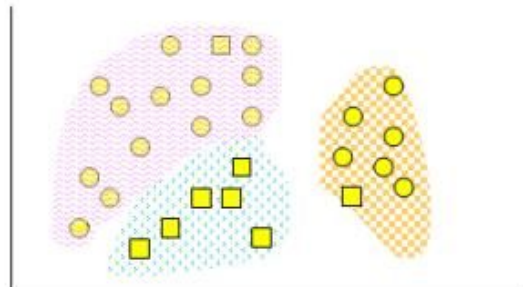
Контролируемое обучение  
Обучение с учителем  
Обучающее множество сопровождается меткой, указывающей класс, к которому относится наблюдение  
Новые данные классифицируются на основании обучающего множества

## Кластеризация

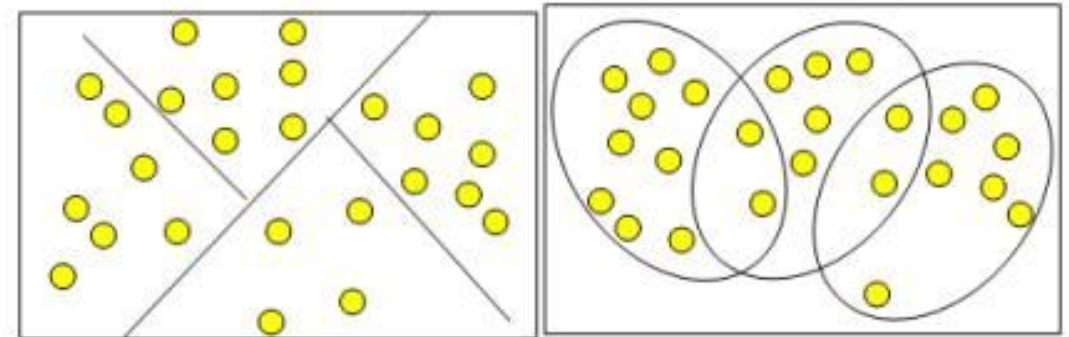
Неконтролируемое обучение  
Обучение без учителя  
Метки класса обучающего множества неизвестны  
Дано множество данных с целью установления существования классов или кластеров данных



*Классификация: классы  
предопределены  
изначально*



*Кластеризация: классы  
не предопределены,  
осуществляется поиск  
наиболее похожих,  
однородных групп*



# Области применения ассоциативных правил

Существуют различные области применения ассоциативных правил:

- - анализ рыночной корзины;
- - представление рекомендованных покупок в интернет-магазинах и т.п.;
- - поиск ошибок в базах данных;
- - медицинская диагностика;
- - анализ белковых последовательностей;
- - анализ данных переписи населения;
- - анализ рождаемости;
- - анализ погодных явлений;
- - анализ показателей жизнедеятельности человека(по фитнес-трекерам и т.п.).

Цель секвенциального анализа - уменьшение необходимого числа наблюдений. Доказано, что в среднем (при систематическом применении) необходимое число наблюдений почти вдвое меньше по сравнению с обычно применяемыми методами работы, при которых необходимое число наблюдений определяют заранее.

**Отличие поиска ассоциативных правил от секвенциального анализа в том, что в первом случае ищется набор объектов в рамках одной транзакции, т.е. такие товары, которые чаще всего покупаются ВМЕСТЕ. В одно время, за одну транзакцию.**