Data mining: задачи, техники и приложения

Николай Пономарев

Математико-механический факультет СПбГУ

2 мая 2024 г.

Немного о терминологии

Две близких области: Information Retrieval (IR) и Data Mining (DM)

Немного о терминологии

Две близких области: Information Retrieval (IR) и Data Mining (DM)

Определение

IR — наука об организации информации и алгоритмах её *быстрого* получения

Немного о терминологии

Две близких области: Information Retrieval (IR) и Data Mining (DM)

Определение

IR — наука об организации информации и алгоритмах её *быстрого* получения

Определение

DM — наука о методах обнаружения в данных ранее *неизвестных* и практически полезных знаний

Немного о терминологии

Две близких области: Information Retrieval (IR) и Data Mining (DM)

Определение

IR — наука об организации информации и алгоритмах её *быстрого* получения

Определение

DM — наука о методах обнаружения в данных ранее *неизвестных* и практически полезных знаний

Формально, наш курс про IR

Задачи I

Определение

Классификация (classification) — процесс разделения новых наблюдений (observation) на предопределенные классы

Определение

Кластеризация (clustering) — процесс разбиения данных (или наблюдений о них) на группы

Задачи II

Определение

Анализ выбросов (outlier analysis) — способ извлечения полезных знаний из выбросов

Определение

Ассоциативный анализ (association analysis) — процесс поиска ассоциаций среди данных, которые удовлетворяют определенным статистическим требованиям

Статистические подходы

Строится статистическая модель, а затем анализируется следующими способами:

Байесовские сети способ выяснения зависимостей между переменными «с помощью» формулы Байеса

Корреляция используется для установления зависимости между фактами

Регрессия установление соответствия между случайными переменными отражающими связь между зависимой переменной и независимыми переменными x

Факторный анализ используется для поиска основных источников корреляции

Машинное обучение и нейронные сети

- Развивает идеи статистических методов
- В каком-то смысле автоматизирует анализ
- Часто результаты лучше
- Модно

Ещё техники

- СУБД
- Генетические алгоритмы
- Нечёткие множества
- Визуализация

Телекоммуникации

Телеком и мобильные операторы используют data mining для

- Маркетинга
  - Классификация и кластеризация ⇒ таргетированная реклама
- Удержания клиентов
  - Классификация и кластеризация ⇒ обнаружение недовольных клиентов
- Создание оптимальных тарифов
- Оптимизация использования инфраструктуры

Продажи

Торговым предприятиям нужен DM для исследования

- Поведения покупателей
  - Ассоциативный анализ
- Корзины
- Выбора товаров
- Расстановки продуктов на полках
  - Кластеризация
- Влияния акций

Медицина

В медицине DM используется для

- Обнаружения и анализа хронических заболеваний
- Поиск эффективных лекарств
- Отслеживать вероятность эпидемий

Зачем ещё это надо?

- Финансы
- Предотвращение преступлений
- Рекомендательные системы
- Реклама

Источник

M. K. Gupta & P. Chandra
A comprehensive survey of data mining
International Journal of Information Technology, 12 (2020) 1243–1257
https://doi.org/10.1007/s41870-020-00427-7