Анализ данных

Андрей Симановский (andrey.simanovsky@hp.com)
Наталья Васильева
Александр Уланов
Сергей Серебряков

О курсе

- Время и место
  - Каждый вторник в 10-00
- Семестр
- Оценивание
  - Допуск к оцениванию: примерно 3 групповых задания в течение семестра
  - Оценивание: зачет или экзамен
- Учебники:
  - Introduction to data mining. *Tan, Steinbach, Kumar*
  - Введение в информационный поиск. *Мэннинг и др.*

Программа

- Введение
- Извлечение информации из текстов
- Извлечение событий
- Кластеризация/классификация
- Анализ изображений
- …

Определения

- Нетривиальный процесс обнаружения новых, потенциально полезных и в конечном счете понятных паттернов в данных
- Процесс извлечения прежде неизвестной, понятной и полезной информации из больших баз данных и использование её для принятия решений
- Набор методов, используемых в процессе извлечения знаний для распознавания ранее неизвестных отношений и паттернов в данных
- Процесс обнаружения полезных паттернов в данных

Отличия от машинного обучения и статистики

- Фокус на приложения (нетривиальный процесс моделирования данных)
- Фокус на большой объем данных (например, кластеризация как средство уменьшения объема вычислений)
- Фокус на интерпретацию результатов

Инструменты

- Weka
- R
- UIMA
- Excel
- Octave(?)

Данные

- Что такое данные? Атрибуты и предсказываемые переменные
- Типы данных:
  - nominal, ordinal, interval, ratio (rational)
- Типы наборов данных:
  - Записи и матрицы, графы, упорядоченные наборы
- Характеристики: размерность, разреженность, точность
- Качество данных (шум, выбросы, отсутствующие, дубликаты)

Предобработка данных

- Агрегация
- Создание выборки
- Уменьшение размерности
- Моделирование набора атрибутов
- Гранулирование
- Преобразование значений атрибутов

Пример анализа данных: ассоциативные правила

- Данные: транзакционные
- Каждая запись / корзина – множество элементов / продуктов

ID	Корзина
1	Хлеб, кола, молоко
2	Пиво, хлеб
3	Пиво, кола, подгузники, молоко
4	Пиво, хлеб, подгузники, молоко
5	Кола, подгузники, молоко

Правила

- (пиво) → (подгузники, молоко)
- (молоко) → (подгузники)
- (подгузники) → (молоко)

Метрики

- Support (маргинальная вероятность):
  - Какая часть данных содержит левую часть правила
- Confidence (условная вероятность)
  - Какая часть данных, содержащих левую часть, влечет правую
- Другие (Lift etc)
- Ищем правила с ограниченным снизу sup и ограниченной снизу conf

Apriori алгоритм

- Цель: сэкономить память, уменьшить вычисления
- Искать правила заданной длины в отдельном проходе
- Идея: sup монотонен относительно вложенности множеств

Ariori идея

- Apriori циклически ищет часто встречающиеся множества возрастающего размера пока новые множества находятся
- Начинает с множеств размера 1
- На каждом шаге дополняет только множества, полученные на предыдущем

Apriori нотация

- L(i) – множество часто встречающихся множеств размера I
- C(i) – множество кандидатов для L(i), полученных из L(i-1)

Apriori пример

- Шаг 1 Сканировать журнал транзакций и определить частоту каждого продукта
- Шаг 2 Вычислить L(1)
- Шаг 3 Сгенерировать C(2)
- Шаг 4 Вычислить L(2)
- Шаг 5 Сгенерировать C(3) применяя свойство монотонности
- Шаг 6 Вычислить L(3)
- Шаг 7 Вычислить C(4), после применения свойства монотонности завершить алгоритм

Apriori: пример

ID	Корзина
1	i1, i2, i5
2	i2, i4
3	i2, i3
4	i1, i2, i4
5	i1, i3
6	i2, i3
7	i1, i3
8	i1, i2, i3, i5
9	i1, i2, i3

MinSup = 2/9; MinConf = 0

Apriori: пример

Частоты отдельных элементов:

Множество	sup
(i1)	6/9
(i2)	7/9
(i3)	6/9
(i4)	2/9
(i5)	2/9

Apriori: пример

- L(1):

Множество	sup
(i1)	6/9
(i2)	7/9
(i3)	6/9
(i4)	2/9
(i5)	2/9

Apriori: пример

- C(2):

Множество	sup
(i1,i2)	4/9
(i1,i3)	4/9
(i1,i4)	1/9
(i1,i5)	2/9
(i2,i3)	4/9
(i2,i4)	2/9
(i2,i5)	2/9
(i3,i4)	0/9
(i3,i5)	1/9
(i4,i5)	0/9

Apriori: пример

- L(2):

Множество	sup
(i1,i2)	4/9
(i1,i3)	4/9
(i1,i5)	2/9
(i2,i3)	4/9
(i2,i4)	2/9
(i2,i5)	2/9

Apriori: пример

- C(3):

Множество	sup
(i1,i2,i3)	2/9
(i1,i2,i5)	2/9

Apriori: пример

- L(3):

Множество	sup
(i1,i2,i3)	2/9
(i1,i2,i5)	2/9

Apriori: пример

- C(4): empty

Apriori: пример

- Ответ (MinConf=0):
  - i1→i2, i2→i1, i1→i3, i3→i1, i1→i5, i5→i1, i2→i3, i3→i2, i2→i4, i4→i2, i2→i5, i5→i2
  - (i1,i2)→i3, i3 →(i1,i2), (i1,i3)→i2, i2 →(i1,i3), (i2,i3)→i1, i1 →(i2,i3), (i1,i2)→i5, i5 →(i1,i2), (i1,i5)→i2, i2 →(i1,i5), (i2,i5)→i1, i1 →(i2,i5)
- Ответ (MinConf=2/3):
  - i1→i2, i1→i3, i3→i1, i5→i1, i3→i2, i2→i3, i4→i2, i5→i2
  - i5 →(i1,i2), (i1,i5)→i2, (i2,i5)→i1