



Санкт-Петербургский
государственный
университет
www.spbu.ru

Анализ данных

Графеева Н.Г.
2018



ПРОБЛЕМА

Современные информационные системы собирают гигантские объемы данных. Сбор данных без последующего глубокого анализа не позволяет использовать максимум имеющейся информации. В результате возникает парадоксальная ситуация – данных много, а пользы от них мало. Только применение полноценной аналитики позволяет трансформировать данные в реальные знания.



Простые методы анализа

- Вычисление разнообразных статистических показателей
- создание специализированных аналитических отчетов
- построение разнообразных графиков и диаграмм
- использование OLAP-инструментов для оперативного вычисления статистики



Глубокий анализ данных

Реальный бизнес характеризуется сложными зависимостями, большими объемами данных, быстрыми изменениями. Технологии глубокого анализа позволяют выявлять в огромных объемах данных нетривиальные закономерности и превращать знания в конкурентные преимущества.



Понятие Data Mining

Data Mining – это процесс обнаружения в больших базах данных нетривиальных и практически полезных закономерностей.



Сравнение формулировок задач OLAP и Data Mining

Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке? (OLAP)

Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками? (Data Mining)



Классы задач Data Mining

- классификация
- кластеризация
- прогнозирование
- поиск ассоциаций
- поиск последовательностей



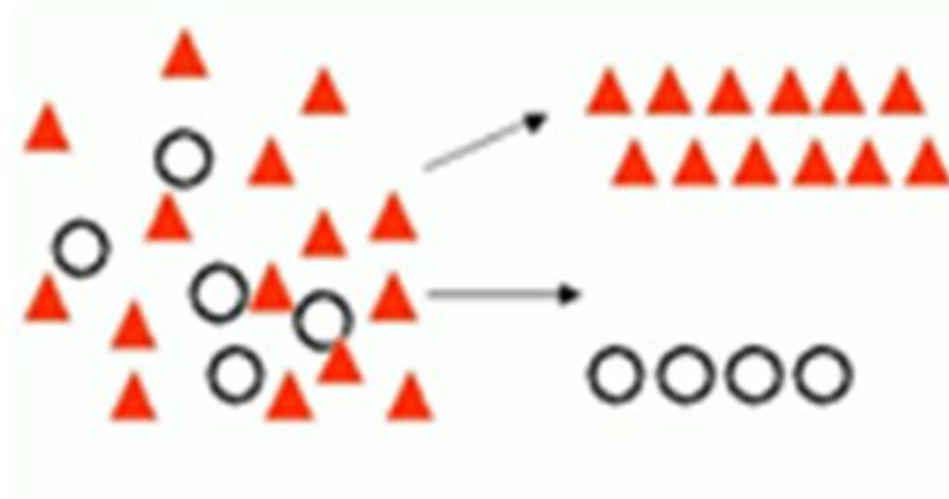
Классификация (Classification)

Задача классификации сводится к определению класса объекта по его характеристикам. В этой задаче множество классов, к которым может быть отнесен объект, известно заранее. Для решения задачи могут использоваться методы:

к-ближайшего соседа (k-Nearest Neighbor);
байесовские сети (Bayesian Networks);
деревья решений; нейронные сети (neural networks) и т.п.



Пример классификации





Кластеризация (Clustering)

Задача кластеризации заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных. Решение этой задачи помогает лучше понять данные. Кроме того, группировка однородных объектов позволяет сократить их число, а следовательно, и облегчить дальнейший анализ.



Пример кластеризации





Живой пример работы алгоритма кластеризации

<http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>

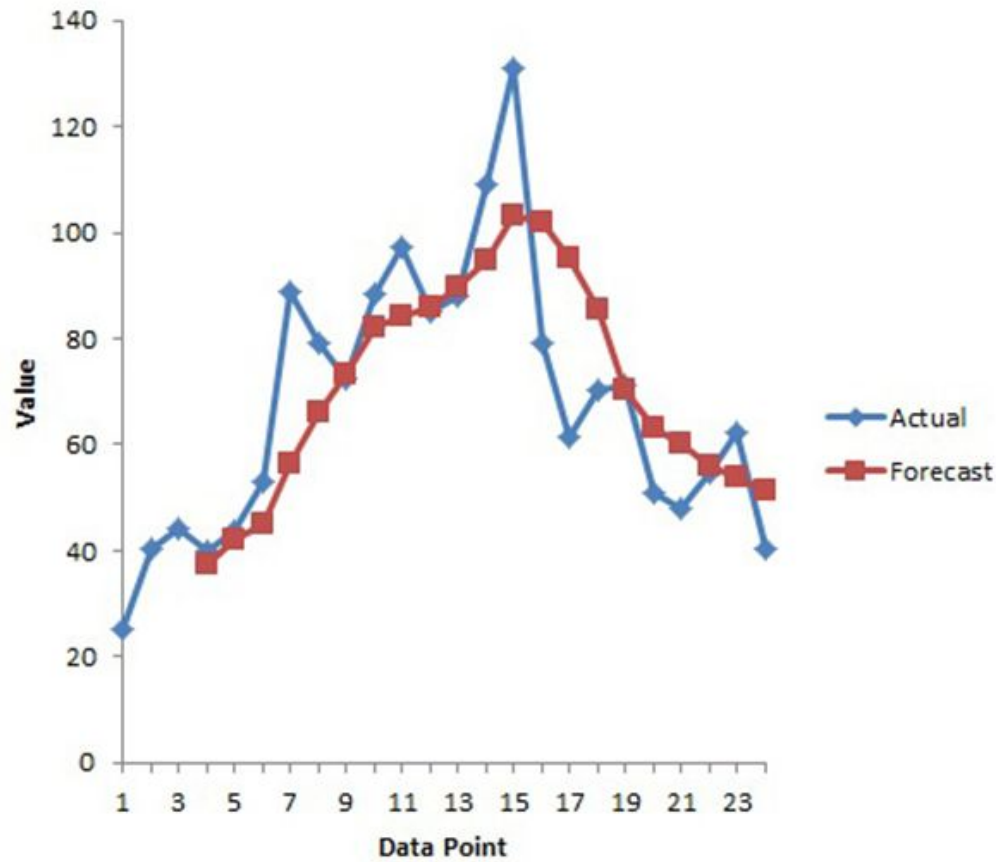


Прогнозирование (Forecasting)

В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей. Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.



Пример прогнозирования



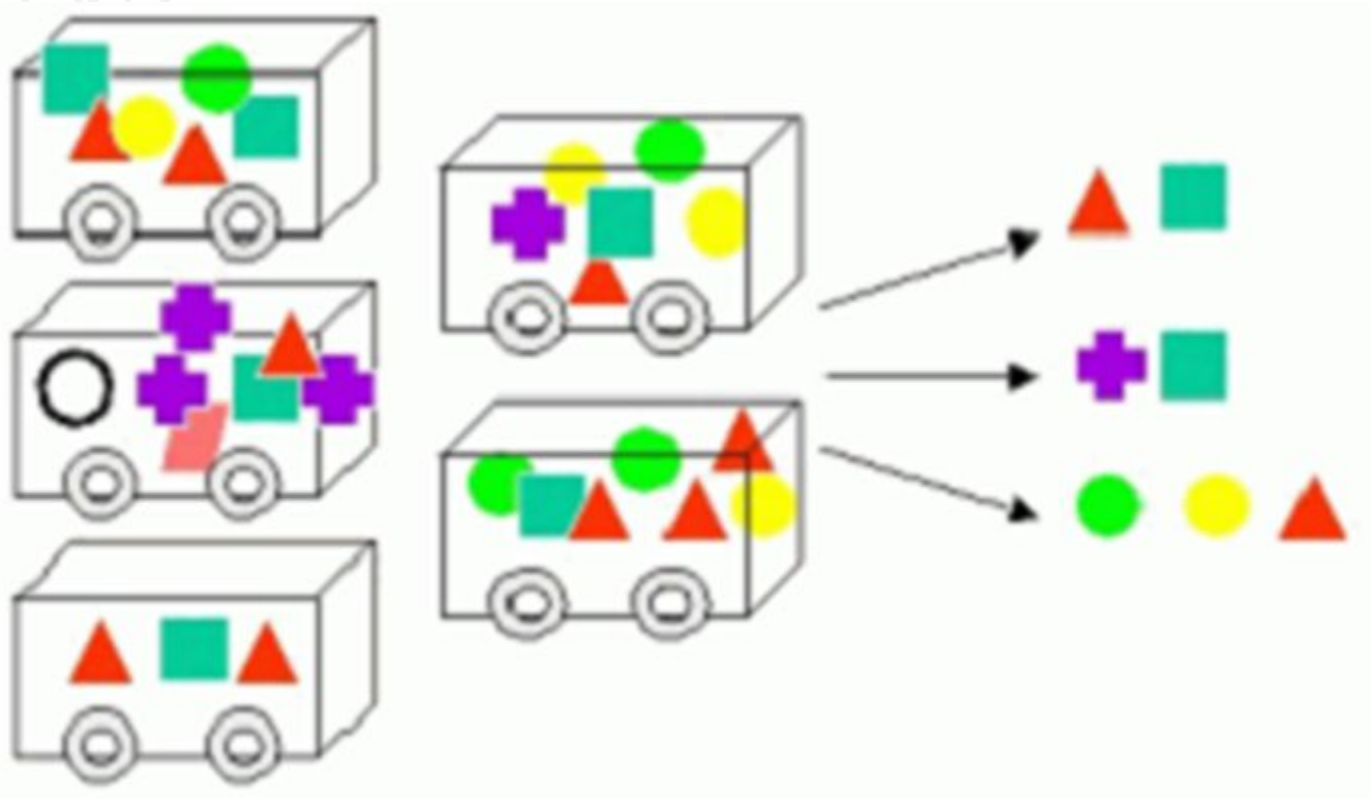


Ассоциации (Associations)

При поиске ассоциативных правил целью является нахождение частых зависимостей между объектами. Найденные зависимости представляются в виде правил и могут быть использованы для лучшего понимания природы анализируемых данных. Наиболее известный алгоритм поиска ассоциативных правил – Apriori.



Пример нахождения ассоциативных правил





Последовательность (Sequence)

Последовательность (sequential association) - временные закономерности между событиями. Последовательность определяется высокой вероятностью цепочки связанных во времени событий. Ассоциация является частным случаем последовательности с временным интервалом, равным нулю. Эту задачу также называют задачей нахождения последовательных шаблонов (sequential pattern).



Сфера применения

Методы Data Mining сегодня интересуют коммерческие предприятия, обладающие большими информационными хранилищами данных. Data Mining представляет большую ценность для руководителей и аналитиков в их повседневной деятельности.



Некоторые бизнес-приложения Data Mining

- розничная торговля
- банковское дело
- телекоммуникации
- страхование
- и другие приложения в бизнесе...



Розничная торговля

- анализ покупательской корзины
- исследование временных шаблонов
- создание прогнозирующих моделей



Банковское дело

- выявление мошенничества с кредитными карточками
- сегментация клиентов
- прогнозирование изменений клиентуры



Телекоммуникации

- выявление категорий клиентов с похожими стереотипами поведения
- выявление лояльности клиентов



Страхование

- выявление мошенничества
- анализ рисков по страховым выплатам



Другие приложения в бизнесе

- поощрение любителей авиаперелетов
- прогнозирование гарантийных обращений к производителям продукции
- развитие автомобильной промышленности с учетом наиболее востребованных опций
- и т.п.



Программные продукты Data Mining

- ❑ аналитические пакеты в некоторых СУБД (например, в ORACLE, DB2, Microsoft SQL Server)
- ❑ библиотеки алгоритмов Data Mining с соответствующей инфраструктурой
- ❑ узкоспециализированные решения



Проблемы существующих решений

Data Mining – бурно развивающаяся мультидисциплинарная отрасль, в которой постоянно появляются новые методы извлечения знаний. Существующие программные продукты либо не успевают, либо не очень следят за такими методами.



Аналитический пакет ORACLE 12

Например, в СУБД ORACLE в 12 версии (выпущена в 2013 году) реализован единственный алгоритм для поиска ассоциативных правил – Apriori (дата публикации – 1994 год). Хотя с тех пор в авторитетных изданиях были опубликованы не менее 11 более совершенных алгоритмов...



Наши работы в области Data Mining

- ❑ Выявление и классификация аномалий магнитного поля с помощью алгоритмов кластеризации (на примере археологических раскопок).
- ❑ Анализ лог-файлов для обнаружения разного рода сбоев в работе аппаратных комплексов.
- ❑ Анализ транспортных потоков Санкт-Петербурга.



Наши работы в области Data Mining

- Прогнозирование потребления продуктов в сети ресторанов.
- Прогнозирование потребления электроэнергии.
- Поиск ассоциативных правил для профилирования ресторанов.
- И многие другие...



Ваши вопросы?

