Этапы анализа данных

Графеева Н.Г.

2016

Последовательность этапов Data Mining

Выдвижение гипотез

1. Максимально использовать знание. экспертов о предметной области.
2. Полагаться на здравый смысл.
3. Отталкиваться от опыта и интуиции специалистов.
4. Собрать и систематизировать максимум возможных предположений и гипотез.

Сбор и систематизация данных (подбор факторов)

1. Абстрагироваться от существующих информационных систем и имеющихся в наличии данных.
2. Описать факторы, влияющие на анализируемый процесс/объект.
3. Оценить значимость каждого фактора.

Сбор и систематизация данных (методы сбора)

1. Получение из существующих информационных систем.
2. Извлечение необходимых сведений из косвенных данных.
3. Использование открытых источников .
4. Проведение социологических, маркетинговых и подобных исследований .
5. Ввод данных «вручную».

Сбор и систематизация данных. Формат.

- Данные должны быть собраны в единую таблицу в формате MS Excel, текстовые файлы с разделителями или в набор таблиц в любой СУБД.

- Необходимо унифицировать представление данных – один и тот же объект должен описываться везде одинаково.

Сбор упорядоченных данных

Задачи **классификации, кластеризации и регрессии**.

В случае анализа временных рядов каждому столбцу соответствует один фактор, а в каждую строку заносятся упорядоченные по времени события с единым интервалом между строками.

NN	Дата	Частота закупок	Объем продаж
1	01.05.04	256	3970
2	02.05.04	278	5870

Объемы упорядоченных данных

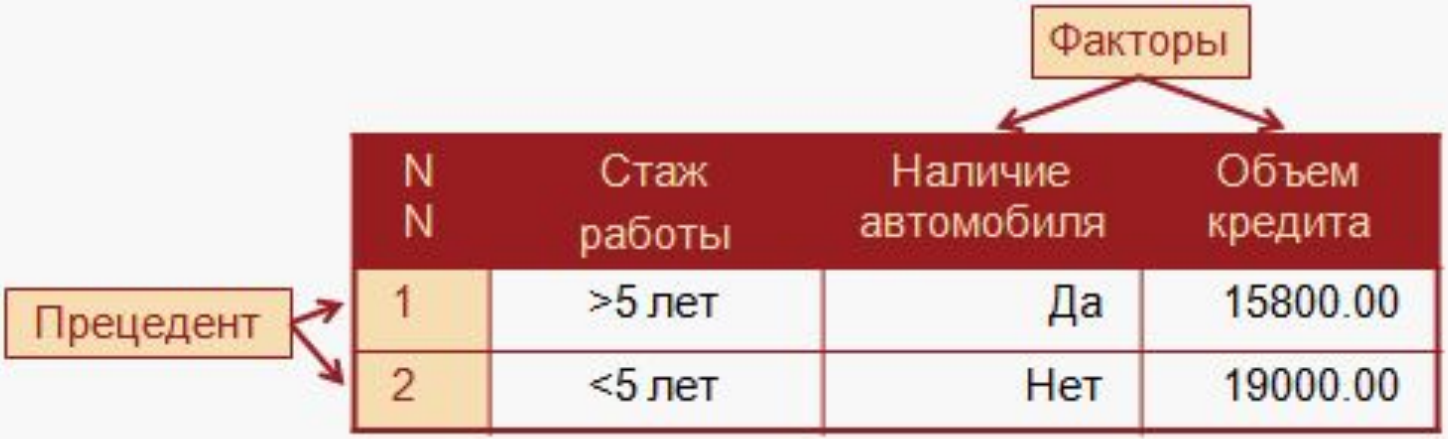
- Если для процесса характерна сезонность/цикличность, необходимо иметь данные хотя бы за один полный сезон/цикл с возможностью варьирования интервалов (понедельное, помесячное…).

- Максимальный горизонт прогнозирования зависит от объема данных:
  - данные на 1,5 года – прогноз максимум на 1 месяц
  - данные за 2-3 года – прогноз максимум на 2 месяца

Сбор неупорядоченных данных

Задачи **классификации, кластеризации и регрессии**.

В случае анализа не связанных по времени событий каждому столбцу соответствует фактор, а в каждую строку заносится пример (ситуация, прецедент). Упорядоченность строк не требуется.



N N	Стаж работы	Наличие автомобиля	Объем кредита
1	>5 лет	Да	15800.00
2	<5 лет	Нет	19000.00

Объемы неупорядоченных данных

1. Количество примеров (прецедентов) должно быть значительно больше количества факторов.
2. Желательно, чтобы данные покрывали как можно больше ситуаций реального процесса.
3. Пропорции различных примеров (прецедентов) должны примерно соответствовать реальному процессу.

Сбор транзакционных данных

Задача **поиска ассоциативных правил**.

Под транзакцией подразумевается несколько объектов или действий, являющихся логически связанной единицей. Очень часто данный механизм используется для анализа покупок (чеков) в супермаркетах. Но в общем случае речь может идти о любых связанных объектах или действиях.

Одна транзакция

Код транзакции	Товар
10200	Йогурт «Чудо» 0,4
10200	Батон «Рязанский»
10201	Вода «Боржоми» 0,5
10201	Сахарный песок

Объемы транзакционных данных

- Анализ транзакций целесообразно производить на большом объеме данных, иначе могут быть выявлены статистически необоснованные правила. Алгоритмы поиска ассоциативных связей способны быстро перерабатывать огромные массивы данных.

- Примерное соотношение между количеством объектов и объемом данных:
  - 300-500 объектов – более 10 тыс. транзакций
  - 500-1000 объектов – более 300 тысяч транзакций

Подбор модели

1. Уделить внимание очистке данных.
2. Комбинировать методики анализа.
3. Не гнаться за абсолютной точностью и начать использование при получении первых приемлемых результатов.
4. При невозможности получения приемлемых результатов вернуться на предыдущие шаги схемы.

Тестирование, интерпретация

1. Для оценки полученных результатов использовать знания экспертов.
2. Тестировать построенные модели на различных выборках для оценки их обобщающих способностей.
3. При невозможности получения приемлемых результатов вернуться на предыдущие шаги схемы.

Использование

1. При получении приемлемых результатов начать использование.
2. Периодически оценивать адекватность модели текущей ситуации. Даже самая удачная модель со временем перестает ей соответствовать.
3. Постоянно работать над улучшением модели.

Задание 0

- Загрузить в базу содержимое следующего файла (понадобится для последующих заданий):

goodsamount.xml