УДК

МЕТОДЫ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ В DATA MINING

Петце А.О.
Донецкий Национальный технический университет,
Факультет компьютерных наук и технологий,
Кафедра автоматизированных систем управления
E-mail: arthurpirojkoff@mail.ru

Аннотация

Петце А.О. Методы предварительной обработки данных в Data Mining. *Рассмотрены методы предварительной обработки данных в условиях интеллектуального анализа данных (Data Mining).*

Общая постановка проблемы.

Data Mining – интеллектуальный анализ данных, на основе существующих данных (БД, Хранилища Данных, OLAP-системы).

Данные – результат фиксации некоторой информации, сами могут выступать как источник «скрытой» информации. Основные требования к такой «скрытой» информации:

- ранее неизвестна;
- не тривиальна;
- практически полезна;
- доступна для интерпретации.

Эти требования определяют суть методов Data Mining, и то какие из них будут использованы в сборе, подготовке, анализе данных. Перед анализом и применением алгоритмов Data Mining, необходимо провести предварительную обработку данных.

Предварительная обработка данных является важнейшим этапом, от качества выполнения которого, зависит возможность получения качественных результатов всего процесса Data Mining. Сама предварительная обработка данных включает два направления: очистка и оптимизация данных. Также, следует помнить, что по некоторым оценкам этап предварительной обработки данных может занять до 80% всего времени, отведенного на проект.

Исследования.

Собственно, предварительная обработка данных позволяет как повысить качество интеллектуального анализа данных, так и повысить качество самих данных.

Один из прогнозов увеличения качества данных, сделанный Даффи Брансоном [4] (Duffy Brunson) звучит следующим образом:

«Прогноз. Многие компании стали обращать больше внимания на качество данных, поскольку низкое качество данных стоит денег в том смысле, что ведет к снижению производительности, принятию неправильных бизнес-решений и невозможности получить желаемый результат, а также затрудняет выполнение требований законодательства. Поэтому компании действительно намерены предпринимать конкретные действия для решения проблем качества данных.

Реальность. Данная тенденция сохраняется, особенно в индустрии финансовых услуг. В первую очередь это относится к фирмам, старающимся выполнять соглашение Basel II. Некачественные данные не могут использоваться в системах оценки рисков, которые применяются для установки цен на кредиты и вычисления потребностей организации в капитале. Интересно отметить, что существенно изменились взгляды на способы решения

проблемы качества данных. Вначале менеджеры обращали основное внимание на инструменты оценки качества, считая, что "собственник" данных должен решать проблему на уровне источника, например, очищая данные и переобучая сотрудников. Но сейчас их взгляды существенно изменились. Понятие качества данных гораздо шире, чем просто их аккуратное введение в систему на первом этапе. Сегодня уже многие понимают, что качество данных должно обеспечиваться процессами извлечения, преобразования и загрузки (extraction, transformation, loading - ETL), а также получения данных из источников, которые подготавливают данные для анализа».

Очистка необходима для повышения качества данных, что в свою очередь повышает скорость и качество анализа данных методами Data Mining. Она включает в себя методы устранения дублирования, противоречивости и недостоверности, восстановления и/или заполнения пробелов, сглаживания и очистки данных от шума.

Оптимизация же удваивает эффект, и влияет на те же характеристики анализа, что и очистка данных. Заключается в методах снижения размерности данных, выявления и исключения незначащих признаков.

Очистка данных.

Анализу поддаются абсолютно все данные, как качественные, так и не очень. Однако именно от качества данных зависит качество анализа, поэтому не качественные данные (так называемые «грязные») подвергаются очистке, дабы нормировать общий набор данных перед анализом.

Грязные данные могут появиться по абсолютно разным причинам, естественно, что все эти причины берут начало от человеческого фактора. Из всех причин, можно выделить следующие:

1) ошибка при вводе данных;
2) использование отличных от остальных форматов представления или единиц измерения;
3) несоответствие стандартам;
4) отсутствие своевременного обновления;
5) неудачное удаление записей дубликатов и т.п.

Однако важно понимать, что средства очистки могут справиться не со всеми видами «грязных» данных.

Основные варианты ошибок, в так называемых «грязных» данных, следующие:

противоречивость информации;
пропуски в данных;
аномальные значения;
шум.

Для решения каждой из этих проблем есть свои методы решения, однако целесообразно будет рассмотреть те, которые не зависят от предметной области и поставленной задачи анализа.

Противоречивость информации

В первую очередь, следует определиться с тем, какую информацию считать противоречивой. После этого, есть несколько вариантов действий:

1) обнаружив некоторое количество противоречивых данных, удалить их; самый простой, однако не самый разумный способ.

2) исправить противоречивые данные (один из вариантов – вычислить вероятность появления каждого из противоречивых событий и выбрать наиболее вероятный); наиболее грамотный способ.

Пропуски в данных

Наиболее актуальная проблема для большинства хранилищ данных. Большая часть методов прогнозирования (одной из причин Data Mining) исходят из предположения, что данные поступают, равномерным потоком. На практике такое встречается крайне редко, т.к. БД хранилищ могут отличаться по структуре, в зависимости от законодательства и других требований. Поэтому, прогнозирование на основе таких данных реализуется некачественно или со значительными ограничениями. Для исправления таких ошибок существуют такие методы как:

1) аппроксимация, - если данные отсутствуют в какой-то точке, берется её окрестность и вычисляется значение в этой точке, добавляя соответствующую запись в хранилище.
2) определение наиболее правдоподобного значения; для этого берутся все данные, а не окрестность точки.

Аномальные значения

В этом случае, аномальными называют данные, которые выбиваются из общей выборки данных. Такие данные лучше откорректировать и привести к наиболее вероятному значению, потому как аномальное значение при прогнозировании, будет считаться абсолютно нормальным. Для аномальных значений применяются следующие действия:

1) значение удаляется.
2) значение заменяется на ближайшее граничное.

Шум

Шум в данных, своего рода отклонения от среднего значения данных, не несет никакой полезной информации, в том числе и «скрытой». Вот некоторые методы борьбы с шумом:

1) спектральный анализ, - с его помощью отсекаются высокочастотные составляющие данных; в общем, это частые и незначительные колебания вокруг основного сигнала.
2) авторегрессионные методы, - применяются при анализе временных рядов, и сводятся к нахождению функции, которая описывает процесс и шум; после этого шум можно будет убрать, оставив основной сигнал.

Оптимизация данных

Предварительное снижение размерности

Более понятные и прозрачные результаты анализа в Data Mining могут быть получены, если из множества исходных переменных и значений использовать некие обобщенные переменные. Таким образом, возникает снижение размерности данных. Задача снижения размерности может решаться различными методами, основные из которых:

1 метод главных компонент, - сокращение размерности пространства переменных и признаков с минимальной потерей полезной информации, компоненты приводятся в виде убывающей регрессии кол-ва признаков в них (сам метод г.к. является разновидностью метода факторного анализа).

2 метод факторного анализа, - изучает и устанавливает взаимосвязи между переменными, преследует две цели: сокращение числа переменных, классификация данных; при помощи ф.а. большое число переменных сводится к меньшему числу независимых величин – факторов.

Выявление и исключение незначащих признаков

Метод поиска и устранения таких признаков, которые наименее взаимосвязаны с выходным результатом. Такие признаки исключаются без потери полезной информации.

Критерием принятия решения об исключении является порог значимости. Если взаимосвязь между признаком и выходным результатом меньше порога значимости, то соответствующий признак отбрасывается как незначащий.

Выводы

На данный момент Data Mining является очень популярным направлением анализа и прогнозирования в целях улучшения производства. Этапы такого анализа позволяют не только сделать прогнозы и принять решения, но также классифицировать, нормировать и повысить качество данных используемых в IT-системах.

Процесс интеллектуального анализа данных, коим является и Data Mining, достаточно трудоемок и долог. До 80% этого процесса может занять этап предварительной обработки данных. Поэтому очевидна насущность данной проблемы, поиска оптимальных и необходимых методов обработки данных перед их непосредственным анализом.

На данный момент, существует достаточное количество инструментов по проведению всех этапов Data Mining, в том числе и предварительной обработки данных. Однако область интеллектуального анализа данных, остается открытой для нововведений, поскольку относительно нова.

Список литературы

1. А.А. Ежов, С.А. Шумский Нейрокомпьютинг и его применение в экономике и бизнесе. – М, 1998. – С. 126-145.

2. Чубукова И.А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – С. 208-221.

3. www.basegroup.ru

4. Десять основных тенденций 2005 года в области Business Intelligence и Хранилищ данных - http://citforum.ru/gazeta/3/

5. Великие раскопки и великие вызовы - http://www.kdnuggets.com/gpspubs/piatetsky-interview-computerra.pdf

6. Чубукова И.А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – С. 208-221.

7. www.basegroup.ru