электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.6

Методы интеллектуального анализа данных

***Гаврилова М.А.**, студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»

Научный руководитель: Ерёмин О.Ю., к.т.н, ассистент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»
Ereminou@bmstu.ru

Введение

Стремительное развитие информационных технологий, в частности, прогресс в методах сбора, хранения и обработки данных позволил многим организациям собирать огромные массивы данных, которые необходимо анализировать. Объемы этих данных настолько велики, что возможностей экспертов уже не хватает.

На сегодняшний день интенсивно развивается направление, связанное с интеллектуализацией методов обработки и анализа данных, другими словами Data Mining.

Понятие Data Mining

Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.

Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, «извлечение зерен знаний из гор данных», раскопка знаний в базах данных, информационная проходка данных, «промывание» данных. Интеллектуальный анализ данных затрагивает такие области как: прикладная статистика, распознование образов, нейронные сети, искусственый интеллект, базы данных и т.д.

Существует множество различных методов интеллектуального анализа данных, моделирования запросов, обработки и сбора информации. Но прежде чем описывать

методы, используемые для анализа Data Mining, рассмотрим интеллектуальный анализ данных как процесс, опираясь на технологию описываемую компанией IBM.

По сути, интеллектуальный анализ данных — это обработка информации и выявление в ней моделей и тенденций, которые помогают принимать решения. Принципы интеллектуального анализа данных известны в течение многих лет, но с появлением больших данных они получили еще более широкое распространение.

Большие данные привели к взрывному росту популярности более широких методов интеллектуального анализа данных, отчасти потому, что информации стало гораздо больше, и она по самой своей природе и содержанию становится более разнообразной и обширной. При работе с большими наборами данных уже недостаточно относительно простой и прямолинейной статистики. Имея 30 или 40 миллионов подробных записей о покупках, недостаточно знать, что два миллиона из них сделаны в одном и том же месте. Чтобы лучше удовлетворить потребности покупателей, необходимо понять, принадлежат ли эти два миллиона к определенной возрастной группе, и знать их средний заработок [1].

Эти бизнес-требования привели от простого поиска и статистического анализа данных к более сложному интеллектуальному анализу данных. Для решения бизнес-задач требуется такой анализ данных, который позволяет построить модель для описания информации и в конечном итоге приводит к созданию результирующего отчета. Этот процесс иллюстрирует рис. 1.

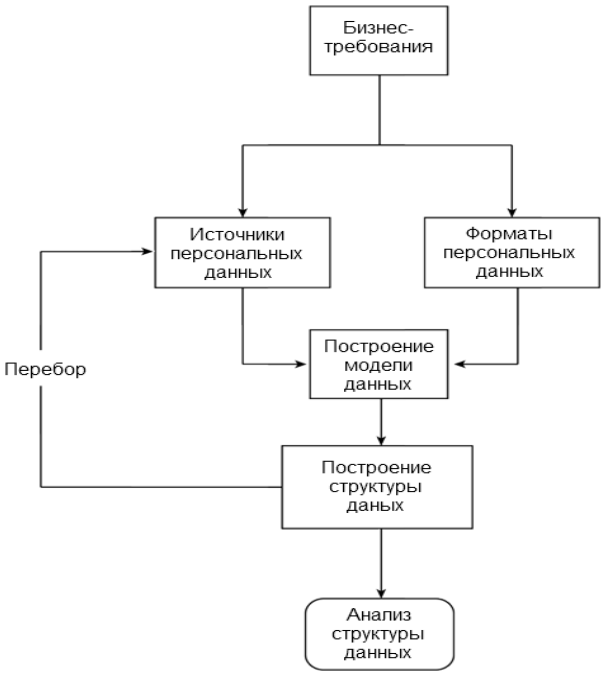

Рис. 1. Схема процесса

Процесс анализа данных, поиска и построения модели часто является итеративным, так как нужно разыскать и выявить различные сведения, которые можно извлечь. Необходимо также понимать, как связать, преобразовать и объединить их с другими данными для получения результата. После обнаружения новых элементов и аспектов данных подход к выявлению источников и форматов данных с последующим сопоставлением этой информации с заданным результатом может измениться.

Большинство аналитических методов, используемые в технологии Data Mining – это известные математические алгоритмы и методы. Новым в их применении является возможность их использования при решении тех или иных конкретных проблем, обусловленная появившимися возможностями технических и программных средств. Следует отметить, что большинство методов Data Mining были разработаны в рамках теории искусственного интеллекта[5].

Метод представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Различные методы Data Mining характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных. Основные свойства и характеристики методов Data Mining: точность, масштабируемость, интерпретируемость, проверяемость, трудоемкость, гибкость, быстрота и популярность.

Масштабируемость – свойство вычислительной системы, которое обеспечивает предсказуемый рост системных характеристик, например, быстроты реакции, общей производительности и пр., при добавлении к ней вычислительных ресурсов.

Сравнительная характеристика распространенных методов представлена в таблице. Оценка характеристик представлена категориями: чрезвычайно низкая, очень низкая, низкая/нейтральная, нейтральная/низкая, нейтральная, нейтральная/высокая, высокая, очень высокая.

Алгоритм / Признак	Линейная регрессия	Нейронные сети	Методы визуализации	Деревья решений	k-ближайшего соседа
Точность	нейтрал.	высокая	высокая	низкая	низкая
Масштабируемость	высокая	низкая	низкая	высокая	очень низкая
Интерпретируемость	высокая/ нейтрал.	низкая	высокая	высокая	высокая/ нейтрал
Проверяемость	высокая	низкая	высокая	высокая/ нейтрал.	нейтрал.
Трудоемкость	нейтрал.	нейтрал.	очень высокая	высокая	нейтрал. низкая
Разносторонность	нейтрал.	низкая	низкая	высокая	низкая
Быстрота	высокая	очень низкая	чрезвычайно низкая	высокая/ нейтрал.	высокая
Популярность	низкая	низкая	высокая/ нейтрал.	высокая/ нейтрал	низкая

Каждый из методов, рассмотренный в таблице 1, имеет свои сильные и слабые стороны. Но ни один метод не может в полной мере обеспечить решение всего спектра задач Data Mining.

В интеллектуальном анализе данных различают две группы методов:

- статистические методы, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;
- кибернетические методы, включающие множество разнородных математических подходов.

Преимуществом такой классификации является ее удобство для интерпретации - она используется при описании математических средств современного подхода к извлечению знаний из массивов исходных наблюдений (оперативных и ретроспективных), т.е. в задачах Data Mining.

Рассмотрим подробнее представленные выше группы.

Статистические методы Data Mining

В эти методы представляют собой четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных;
- выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
- многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, факторный анализ и др.);
- динамические модели и прогноз на основе временных рядов.

Арсенал статистических методов Data Mining классифицирован на четыре группы методов:

1. Дескриптивный анализ и описание исходных данных.
2. Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
4. Анализ временных рядов (динамические модели и прогнозирование)[2,3].

Кибернетические методы Data Mining

Второе направление Data Mining - это множество подходов, объединенных идеей компьютерной математики и использования теории искусственного интеллекта.

К этой группе относятся такие методы: искусственные нейронные сети (распознавание, кластеризация, прогноз); эволюционное программирование; генетические алгоритмы (оптимизация); ассоциативная память (поиск аналогов, прототипов); нечеткая логика; деревья решений; системы обработки экспертных знаний.

Далее рассмотрим некоторые из представленных методов.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных"[2].

Исследуя один или более атрибутов или классов, можно сгруппировать отдельные элементы данных вместе, получая структурированное заключение. На простом уровне при кластеризации используется один или несколько атрибутов в качестве основы для определения кластера сходных результатов. Кластеризация полезна при определении различной информации, потому что она коррелируется с другими примерами, так что можно увидеть, где подобия и диапазоны согласуются между собой.

Метод кластеризации работает в обе стороны. Можно предположить, что в определенной точке имеется кластер, а затем использовать свои критерии идентификации, чтобы проверить это. График, изображенный на рисунке 2, демонстрирует наглядный пример. Здесь возраст покупателя сравнивается со стоимостью покупки. Разумно ожидать, что люди в возрасте от двадцати до тридцати

лет (до вступления в брак и появления детей), а также в 50-60 лет (когда дети покинули дом) имеют более высокий располагаемый доход.

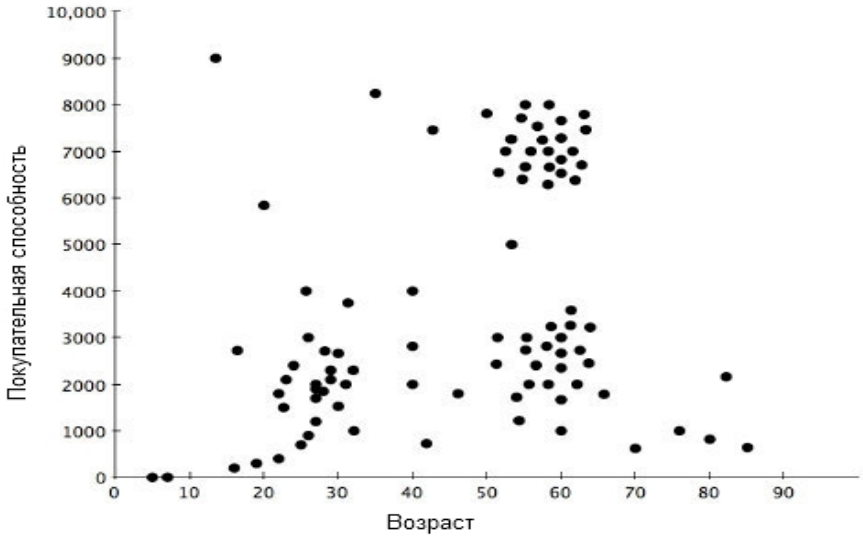

Рис. 2. Кластеризация

В этом примере видны два кластера, один в районе $2000/20-30 лет и другой в районе $7000-8000/50-65 лет. В данном случае мы выдвинули гипотезу и проверили ее на простом графике, который можно построить с помощью любого подходящего ПО для построения графиков. Для более сложных комбинаций требуется полный аналитический пакет, особенно если нужно автоматически основывать решения на информации о ***ближайшем соседе.***

Такое построение кластеров являет собой упрощенный пример так называемого образа ***ближайшего соседа***. Отдельных покупателей можно различать по их буквальной близости друг к другу на графике. Весьма вероятно, что покупатели из одного и того же кластера разделяют и другие общие атрибуты, и это предположение можно использовать для поиска, классификации и других видов анализа членов набора данных.

Метод кластеризации можно применить и в обратную сторону: учитывая определенные входные атрибуты, выявлять различные артефакты. Например, недавнее исследование четырехзначных PIN-кодов выявили кластеры чисел в диапазонах 1-12 и 1-31 для первой и второй пар. Изобразив эти пары на графике, можно увидеть кластеры, связанные с датами (дни рождения, юбилеи)[1,2,3,4].

Метод под названием ***прогнозирование*** хорошо знаком бизнесменам: анализируя данные прошлых периодов, можно построить прогноз на будущее – причем чем подробнее исторические данные и чем больше анализируемый отрезок времени, тем точнее получатся результаты.

Этот метод нередко применяется для оценки спроса на услуги и товары, прогнозирования структуры сбыта, характеризующегося сезонными колебаниями, или потребности в кадрах. Если, к примеру, директор ресторана быстрого питания хочет определить, сколько гамбургеров заказывать на ноябрь, он должен проанализировать цифры ноябрьских продаж в минувшие пять лет[6].

Интеллектуальный анализ данных — это не только выполнение некоторых сложных запросов к данным, хранящимся в базе данных. Независимо от того, используете ли вы SQL, базы данных на основе документов, такие как Hadoop, или простые неструктурированные файлы, необходимо работать с данными, форматировать или реструктурировать их. Требуется определить формат информации, на котором будет основываться ваш метод и анализ. Затем, когда информация находится в нужном формате, можно применять различные методы (по отдельности или в совокупности), не зависящие от требуемой базовой структуры данных или набора данных. Умение оперировать большими данными, применять к ним методы анализа, чтобы в конце концов получить результат – залог успеха любого бизнеса в современном обществе, вступившем в эру безраздельной власти информации.

Список литературы

1. Martin C. Brown. Data mining techniques. Available at: http://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques/, accessed 20.01.2015.
2. Ерёмин О.Ю., Тумковский С.Р. Использование адаптивно-резонансной теории для обнаружения дефектов паяных соединений и повышениякачества печатных плат // Качество. Инновации. Образование. 2010. № 4 (59). С. 37-42.
3. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. Спб: БХВ-Петербург, 2004. 336 с.
4. Обзор методов Data Mining. Режим доступа: http://intellect-tver.ru/?p=165 (дата обращения 20.01.2015).

5. Филиппова Е. Методы интеллектуального анализа данных. Режим доступа: http://datareview.info/article/dannyie-kak-poleznyie-iskopaemyie-osnovnyie-metodyi-data-mining/ (дата обращения 20.02.2015).