Карви Джалал Каис

Научный руководитель: проф.
Брусенцев А.Г
Белгородский государственный технологический университет В. Г.Шухова

АНАЛИЗ ДАННЫХ С ПОМОЩЬЮ DATA MINING

Интеллектуальный анализ данных, или Data Mining, – это процесс обнаружения в недоработанных данных ранее неизвестных, незаурядных, практически полезных и доступных интерпретации знаний, требующихся в различных сферах человеческой деятельности. Современные технологии Data Mining (discovery-driven data mining) обрабатываю сведения с целью автоматического поиска шаблонов (паттернов), свойтсвеенных для каких-либо фрагментов смешанных многомерных данных. В отличие от оперативной аналитической обработки данных (online analytical processing, OLAP) в Data Mining тяготы формулировки гипотез и исследования необычных шаблонов переложено с человека на компьютер.

Известно пять типов задач Data Mining: ассоциация, последовательность, классификация, кластеризация, прогнозирование.

Закономерность типа ассоциация наблюдается в данных, когда несколько события имеют друг с другом тесную связь и происходят при этом в один и тот же момент времени.

Закономерность типа «последовательность» предполагает наличие в данных цепочки связанных друг с другом и распределенных во времени событий.

Закономерность типа «классификация» выделяет в данных на основе исследования признаков уже классифицированных объектов, при этом известна принадлежность объектов к классам. Результатом является формирование правил отнесения объектов к классам.

Закономерность типа «кластеризация» предполагает наличие в данных похожих по каким-либо признакам групп объектов, причем количество групп и принадлежность объектов к ним заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных.

Поиск закономерности типа «прогнозирование» проводится на основе информации, хранящихся в базе данных в виде временных рядов. Если удается создать математическую модель и найти шаблоны,

адекватно показывающие эту динамику, есть вероятность, что с их помощью можно спрогнозировать и как поведет себя система в будущем.

В связи с совершенствованием технологий записи и хранения данных и появлением хранилищ данных на людей рухнули информационные потоки в самых различных отраслях. Стало ясно, что без продуктивной переработки потоки необработанных данных образуют никому не нужную кипу данных.

Особенностью современных требований к обработке данных в хранилищах является то, что данные имеют безграничный объем, данные являются разнообразными. При этом последствия обработки должны быть точны и понятны, а инструменты для обработки необработанных данных должны быть просты для пользователя.

Необходимо отметить, что Data Mining является развитием традиционной математической статистики. Однако методы математической статистики используются для заранее сформулированных суждений.

Исходя из выше изложенного, сделаем вывод, что тема данной работы является актуальной.

1. Понятие анализа данных. Понятие Data Mining

В общем смысле *анализ данных* — это исследования, связанные с многомерной системой данных, имеющей изобилие величин. В процессе исследования данных экспериментатор производит совокупность действий с целью составления определенных понятий о характере явления, описываемого этими данными. Как правило, для анализа данных используются различные математические методы.

Анализ данных невозможно рассматривать только как переработку информации после ее сбора. Анализ данных - это, прежде всего, способ проверки суждений и решения задач исследователя.

Известная оппозиция между познавательными способностями человека, которые ограничены, и бесконечностью Вселенной принуждает нас воспользоваться моделями и моделированием, тем самым облегчает изучение интересующих объектов, явлений и систем.

Построение моделей - универсальный метод изучения окружающего мира, который дает возможность обнаружить зависимости, прогнозировать, разбивать на группы и решать многие другие задачи. Основной смысл моделирования в том, что модель

должна неплохо изображать функционирование моделируемой системы.

Актуальный компьютерный термин Data Mining переводится как «извлечение информации» или «добыча данных». Нередко наряду с Data Mining встречаются термины Knowledge Discovery («обнаружение знаний») и Data Warehouse («хранилище данных»). Появление указанных терминов, которые являются основной частью Data Mining, связано с новым этапом в развитии средств и методов обработки и хранения данных. Итак, цель Data Mining состоит в обнаружении скрытых правил и закономерностей в очень больших объемах данных.

Дело в том, что человеческий интеллект сам по себе не способен воспринимать огромные массивы разнородной информации. В среднем человек не способен воспринять более двух-трех взаимосвязей даже в небольших отрывках. Но статистика так же нередко уступает при решении задач из реальной жизни. Она использует усредненные характеристики отрывки, которых часто являются мнимыми величинами.

Поэтому методы математической статистики оказываются полезными в основном для проверки заранее обоснованных гипотез, тогда как определение гипотезы иногда бывает достаточно сложной и трудоемкой задачей. На сегодняшний день, технологии Data Mining обрабатывают информацию с замыслом автоматического поиска шаблонов (паттернов). В отличие от оперативной аналитической обработки данных (OLAP) в Data Mining бремя формулировки гипотез и выявления необычных (unexpected) шаблонов переложено с человека на компьютер. Data Mining - это совокупность большого числа отличающихся друг от друга методов обнаружения знаний. Выбор метода зависит от типа имеющихся данных и от того, какую информацию нужно получить.

2. Добыча данных - Data Mining

Рассмотрим свойства обнаруживаемых знаний более детально.

Знания должны быть новые, ранее неиспользуемые. Затраченные труды на открытие знаний, о которых пользователь уже имеет понятие, не окупаются. Поэтому важность представляют именно новые, ранее неизведанные знания.

Знания должны быть понятны человеку. Найденные закономерности должны быть объяснимы, в противном случае существует шанс, что они являются случайными.

Знания должны быть полезны на практике. Найденные знания должны быть полезны, особенно на новые данные, а также должны быть точные. Польза от этого заключается в том, чтобы знания могли принести определенную выгоду при их применении.

Знания должны быть необычные. Результаты исследования должны отражать спорные, неожиданные закономерности в данных, составляющие, так называемые скрытые знания. Результаты, которых могли бы быть получены более простым методом, не оправдывают привлечение мощных способов Data Mining.

В Data Mining для понимания полученных знаний служат модели. Классификация моделей зависит от методов их создания. Самыми распространенными являются: правила, деревья решений, кластеры и математические функции.

Сфера применения Data Mining ничем не имеет ограничений - Data Mining используется везде, где есть какие-либо данные. Опыт множества предприятий показывает, что отдача от использования Data Mining может достигать до 1000%. Data Mining имеют большую ценность для руководителей и аналитиков в их деятельности. Бизнесмены осознали, что с помощью методов Data Mining они могут получить значительные преимущества в конкурентной борьбе.

В основу современной технологии Data Mining (discovery-driven data mining) положена идея шаблонов (паттернов), отражающих части разноаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные *подвыборкам данных*, которые могут быть изображены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками предназначенных предположений о устройстве выборке и виде распределений значений исследуемых показателей. Примеры заданий на поиск при использовании Data Mining приведены в табл. 1.

Табл. № 1.

Примеры формулировок задач при использовании методов OLAP и Data Mining

OLAP	**Data Mining**
Каковы средние показатели травматизма для курящих и некурящих?	**Встречаются ли точные шаблоны в описаниях людей, подверженных повышенному травматизму?**

Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов отказавшихся от услуг телефонной компании)?	**Имеются ли характерные портреты клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?**
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	**Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками?**

Важное положение Data Mining — незаурядность разыскиваемых

Рис.1. Уровни знаний, извлекаемых из данных

шаблонов. Это означает, что найденные шаблоны должны отображать неочевидные, неожиданные систематичности в данных, составляющие, так называемые скрытые знания. К обществу пришло понимание, что необработанные данные содержат глубокий пласт знаний, при грамотной раскопке которого могут быть найдены настоящие «сокровища» (рис.1).

Определение термина Data Mining дал в 1996 г. один из основателей этого направления - Григорий Пятецкий-Шапиро.

Data Mining - это процесс обнаружения в необработанных данных ранее неизвестных, незаурядных, практически полезных и доступных к интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

3. Практическое применение Data Mining

Следует отметить, что использование системы Data Mining не исключает применения статистических инструментов и OLAP-средств, так как результаты обработки данных с помощью последних, как правило, содействуют лучшему осознанию характера закономерностей, которые подобает искать.

Использование Data Mining доказано наличием большого количества данных. Данные в хранилище представляют собой постоянно обновляющийся набор, общий для всего предприятия и дает возможность восстановить картину его действий в тот или иной момент времени, а структура данных хранилища проектируется таким образом, чтобы исполнение запросов к нему производилось максимально эффективно. Однако существуют средства Data Mining, способные исполнять поиск закономерностей, зависимости и замысла не только в хранилищах данных, но и в OLAP-кубах, то есть в наборах заранее обработанных статистических данных.

Эксперты считают, что в ближайшее время Data Mining станет одним из наиважнейших направлений разработки программного обеспечения. За счет выявления насыщенной структуры собранной информации и анализа в режиме реальнего времени данная технология будет являться ключевым методом разработки «индивидуальной Сети», предназначенной под определенные нужды каждого пользователя.

Ниже представлены типичные задачи, которые можно решатся с помощью Data Minin в различных сферах деятельности.

Розничная торговля:

- исследование покупательской корзины, рассчитано на выявление товаров, которые покупатели стараются приобретать вместе. Знание покупательской корзины требуется для улучшения рекламы, выработки стратегии создания запасов товаров и способов их размещения в торговых залах;
- анализ временных шаблонов помогает торговым предприятиям принимать решения о создании товарных запасов.
- создание прогнозирующих моделей дает вероятность торговым предприятиям узнать характер потребностей различных слоев населения с определенным поведением. Эти знания предназначены для разработки экономичных мероприятий по продвижению товаров.

Банковское дело:

- обнаружение мошенничества с кредитными карточками.

- сегментация клиентов. Классифицируя клиентов на отдельные категории, банки делают свою маркетинговую политику более целенаправленной и эффективной, предоставляя различные виды услуг разным группам клиентов.
- прогнозирование изменений клиентуры. Data Mining содействует банкам строить прогнозные модели важности своих клиентов и соответствующим образом обслуживать каждую категорию.

Телекоммуникации:

- анализ записей о развернутых характеристиках вызовов. Присутствие такого анализа – открытие категорий клиентов со схожими стереотипами пользования их услугами и разработка интересных наборов цен и услуг;
- обнаружение лояльности клиентов. Data Mining можно использовать для определения характеристик клиентов, которые, один раз обратились к услугами данной компании, вероятнее всего останутся ей верными. В итоге средства, которые выделяются маркетинг, можно тратить там, где выручка будет больше всего.

Страхование:

- разоблачение мошенников. Страховые компании снижают уровень мошенничества, находя определенные стереотипы в заявлениях о выплате страхового возмещения, которые характеризуются взаимоотношением между юристами, врачами и заявителями;
- анализ риска. Путем определения сочетаний факторов, связанных с оплатой заявлений, страховщики могут снизить свои потери по обязательствам.

На основании выше изложенного мы можем сделать ряд выводов:

1. Рынок систем Data Mining развивается. В этом развитии принимают участие все крупнейшие компании. К примеру, Microsoft непосредственно руководит большим сегментом данного рынка.

2. Системы Data Mining применяются по двум направлениям:

1) как валовый продукт для бизнес-приложений;

2) как средство для проведения уникальных исследований. Сегодня стоимость валового продукта от $1000 до $10000. Число инсталляций массовых продуктов на сегодняшний день насчитывает десятки тысяч. Лидеры Data Mining связывают развитие этих систем с применением их как интеллектуальных приложений.

3. Вопреки обилию методов Data Mining, приоритет сдвигается в сторону логических алгоритмов поиска в данных if-then правил. С их помощью можно решать задачи прогнозирования, систематизирования,

распознавания образов, сегментации баз данных, извлечения из данных “скрытой” информации, интерпретации данных, установления ассоциаций в базы данных и др. Результаты таких алгоритмов эффективны и легко объясняются.

4. Основной проблемой логических методов выявления закономерностей является проблема подбора вариантов за продолжительное время. Другие трудности связаны с тем, что известные методы поиска не поддерживают функцию обобщения найденных правил и функцию поиска наилучшей структуры. Решение этих проблем может образовать предмет конкурентоспособных разработок.

Библиографический список :

1. Айвазян С. А., Бухштабер В. М., Юнюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989.
2. Гик Дж., ван. Прикладная общая теория систем. — М.: Мир, 1981.
3. Дюк В.А. Обработка данных на ПК в примерах. — СПб: Питер, 1997.
4. Кречетов Н. Продукты для интеллектуального анализа данных. — Рынок программных средств, № 14–15, 1997, с. 32–39.
5. Киселев М., Соломатин Е.. Средства добычи знаний в бизнесе и финансах. — Открытые системы, № 4, 1997, с. 41–44.
6. Ольга Горчинская Семинары по технологиям Oracle9i. Инструментальные средства Oracle Data Mining http://www.oracle.com
7. Сайт компании SAS www.sas.com
8. Сайт компании StatSoft www.StatSoft.com
9. Сайт www.spc-consulting.ru/dms/