Министерство образования Республики Беларусь

**Витебский государственный технологический университет**

**Контрольная работа**

по дисциплине "КИТ"

Студента группы КДс 10

Сухих Михаила Александровича

№ зач. книжки 08887

Работа выполнена 13 октября 2009

**Витебск 2009**

## 1. Что такое Data Mining?

Data Mining переводится как "добыча" или "раскопка данных". Нередко рядом с Data Mining встречаются слова "обнаружение знаний в базах данных" (knowledge discovery in databases) и "интеллектуальный анализ данных". Их можно считать синонимами Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных.

До начала 90-х годов, казалось, не было особой нужды переосмысливать ситуацию в этой области. Все шло своим чередом в рамках направления, называемого прикладной статистикой (см. например, [1]). Теоретики проводили конференции и семинары, писали внушительные статьи и монографии, изобиловавшие аналитическими выкладками.

Вместе с тем, практики всегда знали, что попытки применить теоретические экзерсисы для решения реальных задач в большинстве случаев оказываются бесплодными. Но на озабоченность практиков до поры до времени можно было не обращать особого внимания - они решали главным образом свои частные проблемы обработки небольших локальных баз данных.

И вот прозвенел звонок. В связи с совершенствованием технологий записи и хранения данных на людей обрушились колоссальные потоки информационной руды в самых различных областях. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) теперь сопровождается регистрацией и записью всех подробностей его деятельности. Что делать с этой информацией? Стало ясно, что без продуктивной переработки потоки сырых данных образуют никому не нужную свалку.

Специфика современных требований к такой переработке следующие:

Данные имеют неограниченный объем

Данные являются разнородными (количественными, качественными, текстовыми)

Результаты должны быть конкретны и понятны

Инструменты для обработки сырых данных должны быть просты в использовании

Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших проблем. Главная причина - *концепция усреднения по выборке*, приводящая к операциям над фиктивными величинами (типа средней температуры пациентов по больнице, средней высоты дома на улице, состоящей из дворцов и лачуг и т.п.). Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез (verification-driven data mining) и для "грубого" разведочного анализа, составляющего основу оперативной аналитической обработки данных (online analytical processing, OLAP).

В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих *фрагменты* многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные *подвыборкам данных*, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборке и виде распределений значений анализируемых показателей. Примеры заданий на такой поиск при использовании Data Mining приведены в табл.1.

**Таблица 1. Примеры формулировок задач при использовании методов OLAP и Data Mining [2]**

|  |  |
| --- | --- |
| OLAP | Data Mining |
| Каковы средние показатели травматизма для курящих и некурящих? | Какие факторы лучше всего предсказывают несчастные случаи? |
| Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов (отказавшихся от услуг телефонной компании)? | Какие характеристики отличают клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании? |
| Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке? | Какие схемы покупок характерны для мошенничества с кредитными карточками? |

Важное положение Data Mining - нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные (unexpected) регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge). К обществу пришло понимание, что сырые данные (raw data) содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки (рис.1).



**Рисунок 1. Уровни знаний, извлекаемых из данных**

В целом технологию Data Mining достаточно точно определяет Григорий Пиатецкий-Шапиро - один из основателей этого направления:

**Data Mining - это процесс обнаружения в сырых данных**

**ранее неизвестных**

**нетривиальных**

**практически полезных**

**и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.**

## 2. Кому это нужно?

Сфера применения Data Mining ничем не ограничена - она везде, где имеются какие-либо данные. Но в первую очередь методы Data Mining сегодня, мягко говоря, заинтриговали коммерческие предприятия, развертывающие проекты на основе информационных хранилищ данных (Data Warehousing). Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Например, известны сообщения об экономическом эффекте, в 10-70 раз превысившем первоначальные затраты от 350 до 750 тыс. дол. [3]. Известны сведения о проекте в 20 млн. дол., который окупился всего за 4 месяца. Другой пример - годовая экономия 700 тыс. дол. за счет внедрения Data Mining в сети универсамов в Великобритании.

Data Mining представляют большую ценность для руководителей и аналитиков в их повседневной деятельности. Деловые люди осознали, что с помощью методов Data Mining они могут получить ощутимые преимущества в конкурентной борьбе. Кратко охарактеризуем некоторые возможные бизнес-приложения Data Mining [2].

## 2.1 Некоторые бизнес-приложения Data Mining

Розничная торговля.

Предприятия розничной торговли сегодня собирают подробную информацию о каждой отдельной покупке, используя кредитные карточки с маркой магазина и компьютеризованные системы контроля. Вот типичные задачи, которые можно решать с помощью Data Mining в сфере розничной торговли:

*анализ покупательской корзины* (анализ сходства) предназначен для выявления товаров, которые покупатели стремятся приобретать вместе. Знание покупательской корзины необходимо для улучшения рекламы, выработки стратегии создания запасов товаров и способов их раскладки в торговых залах.

*исследование временных шаблонов* помогает торговым предприятиям принимать решения о создании товарных запасов. Оно дает ответы на вопросы типа "Если сегодня покупатель приобрел видеокамеру, то через какое время он вероятнее всего купит новые батарейки и пленку?"

*создание прогнозирующих моделей* дает возможность торговым предприятиям узнавать характер потребностей различных категорий клиентов с определенным поведением, например, покупающих товары известных дизайнеров или посещающих распродажи. Эти знания нужны для разработки точно направленных, экономичных мероприятий по продвижению товаров.

Банковское дело.

Достижения технологии Data Mining используются в банковском деле для решения следующих распространенных задач:

*выявление мошенничества с кредитными карточками*. Путем анализа прошлых транзакций, которые впоследствии оказались мошенническими, банк выявляет некоторые стереотипы такого мошенничества.

*сегментация клиентов*. Разбивая клиентов на различные категории, банки делают свою маркетинговую политику более целенаправленной и результативной, предлагая различные виды услуг разным группам клиентов.

*прогнозирование изменений клиентуры*. Data Mining помогает банкам строить прогнозные модели ценности своих клиентов, и соответствующим образом обслуживать каждую категорию.

Телекоммуникации.

В области телекоммуникаций методы Data Mining помогают компаниям более энергично продвигать свои программы маркетинга и ценообразования, чтобы удерживать существующих клиентов и привлекать новых. Среди типичных мероприятий отметим следующие:

*анализ записей о подробных характеристиках вызовов.* Назначение такого анализа - выявление категорий клиентов с похожими стереотипами пользования их услугами и разработка привлекательных наборов цен и услуг;

*выявление лояльности клиентов*. Data Mining можно использовать для определения характеристик клиентов, которые, один раз воспользовавшись услугами данной компании, с большой долей вероятности останутся ей верными. В итоге средства, выделяемые на маркетинг, можно тратить там, где отдача больше всего.

Страхование.

Страховые компании в течение ряда лет накапливают большие объемы данных. Здесь обширное поле деятельности для методов Data Mining:

*выявление мошенничества*. Страховые компании могут снизить уровень мошенничества, отыскивая определенные стереотипы в заявлениях о выплате страхового возмещения, характеризующих взаимоотношения между юристами, врачами и заявителями.

*анализ риска*. Путем выявления сочетаний факторов, связанных с оплаченными заявлениями, страховщики могут уменьшить свои потери по обязательствам. Известен случай, когда в США крупная страховая компания обнаружила, что суммы, выплаченные по заявлениям людей, состоящих в браке, вдвое превышает суммы по заявлениям одиноких людей. Компания отреагировала на это новое знание пересмотром своей общей политики предоставления скидок семейным клиентам.

Другие приложения в бизнесе.

Data Mining может применяться во множестве других областей:

*развитие автомобильной промышленности*. При сборке автомобилей производители должны учитывать требования каждого отдельного клиента, поэтому им нужны возможность прогнозирования популярности определенных характеристик и знание того, какие характеристики обычно заказываются вместе;

*политика гарантий*. Производителям нужно предсказывать число клиентов, которые подадут гарантийные заявки, и среднюю стоимость заявок;

*поощрение часто летающих клиентов*. Авиакомпании могут обнаружить группу клиентов, которых данными поощрительными мерами можно побудить летать больше. Например, одна авиакомпания обнаружила категорию клиентов, которые совершали много полетов на короткие расстояния, не накапливая достаточно миль для вступления в их клубы, поэтому она таким образом изменила правила приема в клуб, чтобы поощрять число полетов так же, как и мили.

## 2.2 Специальные приложения

Медицина.

Известно много экспертных систем для постановки медицинских диагнозов. Они построены главным образом на основе правил, описывающих сочетания различных симптомов различных заболеваний. С помощью таких правил узнают не только, чем болен пациент, но и как нужно его лечить. Правила помогают выбирать средства медикаментозного воздействия, определять показания - противопоказания, ориентироваться в лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать исходы назначенного курса лечения и т.п. Технологии Data Mining позволяют обнаруживать в медицинских данных шаблоны, составляющие основу указанных правил.

Молекулярная генетика и генная инженерия.

Пожалуй, наиболее остро и вместе с тем четко задача обнаружения закономерностей в экспериментальных данных стоит в молекулярной генетике и генной инженерии. Здесь она формулируется как определение так называемых маркеров, под которыми понимают генетические коды, контролирующие те или иные фенотипические признаки живого организма. Такие коды могут содержать сотни, тысячи и более связанных элементов.

На развитие генетических исследований выделяются большие средства. В последнее время в данной области возник особый интерес к применению методов Data Mining. Известно несколько крупных фирм, специализирующихся на применении этих методов для расшифровки генома человека и растений.

Прикладная химия.

Методы Data Mining находят широкое применение в прикладной химии (органической и неорганической). Здесь нередко возникает вопрос о выяснении особенностей химического строения тех или иных соединений, определяющих их свойства. Особенно актуальна такая задача при анализе сложных химических соединений, описание которых включает сотни и тысячи структурных элементов и их связей.

Можно привести еще много примеров различных областей знания, где методы Data Mining играют ведущую роль. Особенность этих областей заключается в их сложной системной организации. Они относятся главным образом к надкибернетическому уровню организации систем [4], закономерности которого не могут быть достаточно точно описаны на языке статистических или иных аналитических математических моделей [5]. Данные в указанных областях неоднородны, гетерогенны, нестационарны и часто отличаются высокой размерностью.

## 3. Типы закономерностей

Выделяют пять стандартных типов закономерностей, которые позволяют выявлять методы Data Mining: ассоциация, последовательность, классификация, кластеризация и прогнозирование (рис.2).



**Рисунок 2. Типы закономерностей, выявляемых методами Data Mining**

*Ассоциация* имеет место в том случае, если несколько событий связаны друг с другом. Например, исследование, проведенное в супермаркете, может показать, что 65% купивших кукурузные чипсы берут также и "кока-колу", а при наличии скидки за такой комплект "колу" приобретают в 85% случаев. Располагая сведениями о подобной ассоциации, менеджерам легко оценить, насколько действенна предоставляемая скидка.

Если существует цепочка связанных во времени событий, то говорят о *последовательности*. Так, например, после покупки дома в 45% случаев в течение месяца приобретается и новая кухонная плита, а в пределах двух недель 60% новоселов обзаводятся холодильником.

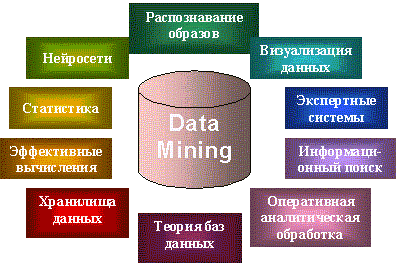
С помощью *классификации* выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил.

*Кластеризация* отличается от классификации тем, что сами группы заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных.

Основой для всевозможных систем *прогнозирования* служит историческая информация, хранящаяся в БД в виде временных рядов. Если удается построить найти шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем.

## 4. Классы систем Data Mining

Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. (рис.3). Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining. Многие из таких систем интегрируют в себе сразу несколько подходов. Тем не менее, как правило, в каждой системе имеется какая-то ключевая компонента, на которую делается главная ставка. Ниже приводится классификация указанных ключевых компонент на основе работы [6]. Выделенным классам дается краткая характеристика.



**Рисунок 3. Data Mining - мультидисциплинарная область**



**Рисунок 4. Популярные продукты для Data Mining**

## 4.1 Предметно-ориентированные аналитические системы

Предметно-ориентированные аналитические системы очень разнообразны. Наиболее широкий подкласс таких систем, получивший распространение в области исследования финансовых рынков, носит название "технический анализ". Он представляет собой совокупность нескольких десятков методов прогноза динамики цен и выбора оптимальной структуры инвестиционного портфеля, основанных на различных эмпирических моделях динамики рынка. Эти методы часто используют несложный статистический аппарат, но максимально учитывают сложившуюся своей области специфику (профессиональный язык, системы различных индексов и пр). На рынке имеется множество программ этого класса. Как правило, они довольно дешевы (обычно $300-1000).

## 4.2 Статистические пакеты

Последние версии почти всех известных статистических пакетов включают наряду с традиционными статистическими методами также элементы Data Mining. Но основное внимание в них уделяется все же классическим методикам - корреляционному, регрессионному, факторному анализу и другим. Самый свежий детальный обзор пакетов для статистического анализа приведен на страницах ЦЭМИ http://is1. cemi. rssi.ru/ruswin/index. htm. Недостатком систем этого класса считают требование к специальной подготовке пользователя. Также отмечают, что мощные современные статистические пакеты являются слишком "тяжеловесными" для массового применения в финансах и бизнесе. К тому же часто эти системы весьма дороги - от $1000 до $15000.

Есть еще более серьезный принципиальный недостаток статистических пакетов, ограничивающий их применение в Data Mining. Большинство методов, входящих в состав пакетов опираются на статистическую парадигму, в которой главными фигурантами служат усредненные характеристики выборки. А эти характеристики, как указывалось выше, при исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами.

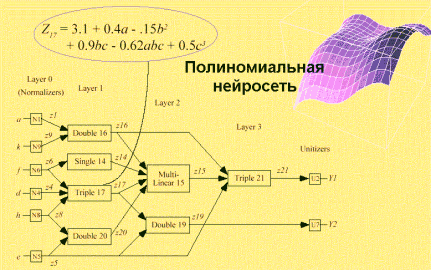
В качестве примеров наиболее мощных и распространенных статистических пакетов можно назвать SAS (компания SAS Institute), SPSS (SPSS), STATGRAPICS (Manugistics), STATISTICA, STADIA и другие.

## 4.3 Нейронные сети

Это большой класс систем, архитектура которых имеет аналогию (как теперь известно, довольно слабую) с построением нервной ткани из нейронов. В одной из наиболее распространенных архитектур, многослойном перцептроне с обратным распространением ошибки, имитируется работа нейронов в составе иерархической сети, где каждый нейрон более высокого уровня соединен своими входами с выходами нейронов нижележащего слоя. На нейроны самого нижнего слоя подаются значения входных параметров, на основе которых нужно принимать какие-то решения, прогнозировать развитие ситуации и т.д. Эти значения рассматриваются как сигналы, передающиеся в следующий слой, ослабляясь или усиливаясь в зависимости от числовых значений (весов), приписываемых межнейронным связям. В результате на выходе нейрона самого верхнего слоя вырабатывается некоторое значение, которое рассматривается как ответ - реакция всей сети на введенные значения входных параметров. Для того чтобы сеть можно было применять в дальнейшем, ее прежде надо "натренировать" на полученных ранее данных, для которых известны и значения входных параметров, и правильные ответы на них. Тренировка состоит в подборе весов межнейронных связей, обеспечивающих наибольшую близость ответов сети к известным правильным ответам.

Основным недостатком нейросетевой парадигмы является необходимость иметь очень большой объем обучающей выборки. Другой существенный недостаток заключается в том, что даже натренированная нейронная сеть представляет собой черный ящик. Знания, зафиксированные как веса нескольких сотен межнейронных связей, совершенно не поддаются анализу и интерпретации человеком (известные попытки дать интерпретацию структуре настроенной нейросети выглядят неубедительными - система "KINOsuite-PR").

Примеры нейросетевых систем - BrainMaker (CSS), NeuroShell (Ward Systems Group), OWL (HyperLogic). Стоимость их довольно значительна: $1500-8000.



**Рисунок 5. Полиномиальная нейросеть**

## 4.4 Системы рассуждений на основе аналогичных случаев

Идея систем case based reasoning - CBR - на первый взгляд крайне проста. Для того чтобы сделать прогноз на будущее или выбрать правильное решение, эти системы находят в прошлом близкие аналоги наличной ситуации и выбирают тот же ответ, который был для них правильным. Поэтому этот метод еще называют методом "ближайшего соседа" (nearest neighbour). В последнее время распространение получил также термин memory based reasoning, который акцентирует внимание, что решение принимается на основании всей информации, накопленной в памяти.

Системы CBR показывают неплохие результаты в самых разнообразных задачах. Главным их минусом считают то, что они вообще не создают каких-либо моделей или правил, обобщающих предыдущий опыт, - в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов CBR системы строят свои ответы.

Другой минус заключается в произволе, который допускают системы CBR при выборе меры "близости". От этой меры самым решительным образом зависит объем множества прецедентов, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза [7].

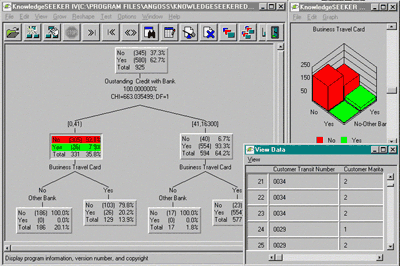
Примеры систем, использующих CBR, - KATE tools (Acknosoft, Франция), Pattern Recognition Workbench (Unica, США).

## 4.5 Деревья решений (decision trees)

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. Они создают иерархическую структуру классифицирующих правил типа "ЕСЛИ... ТО..." (if-then), имеющую вид дерева. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид "значение параметра A больше x?". Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный - то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

Популярность подхода связана как бы с наглядностью и понятностью. Но деревья решений принципиально не способны находить "лучшие" (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и "цепляют" фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода.

Вместе с тем, большинство систем используют именно этот метод. Самыми известными являются See5/С5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада). Стоимость этих систем варьируется от 1 до 10 тыс. долл.



**Рисунок 6. Система KnowledgeSeeker обрабатывает банковскую информацию**

## 4.6 Эволюционное программирование

Проиллюстрируем современное состояние данного подхода на примере системы PolyAnalyst - отечественной разработке, получившей сегодня общее признание на рынке Data Mining. В данной системе гипотезы о виде зависимости целевой переменной от других переменных формулируются в виде программ на некотором внутреннем языке программирования. Процесс построения программ строится как эволюция в мире программ (этим подход немного похож на генетические алгоритмы). Когда система находит программу, более или менее удовлетворительно выражающую искомую зависимость, она начинает вносить в нее небольшие модификации и отбирает среди построенных дочерних программ те, которые повышают точность. Таким образом система "выращивает" несколько генетических линий программ, которые конкурируют между собой в точности выражения искомой зависимости. Специальный модуль системы PolyAnalyst переводит найденные зависимости с внутреннего языка системы на понятный пользователю язык (математические формулы, таблицы и пр).

Другое направление эволюционного программирования связано с поиском зависимости целевых переменных от остальных в форме функций какого-то определенного вида. Например, в одном из наиболее удачных алгоритмов этого типа - методе группового учета аргументов (МГУА) зависимость ищут в форме полиномов. В настоящее время из продающихся в России систем МГУА реализован в системе NeuroShell компании Ward Systems Group.

Стоимость систем до $ 5000.

## 4.7 Генетические алгоритмы

Data Mining не основная область применения генетических алгоритмов. Их нужно рассматривать скорее как мощное средство решения разнообразных комбинаторных задач и задач оптимизации. Тем не менее генетические алгоритмы вошли сейчас в стандартный инструментарий методов Data Mining, поэтому они и включены в данный обзор.

Первый шаг при построении генетических алгоритмов - это кодировка исходных логических закономерностей в базе данных, которые именуют хромосомами, а весь набор таких закономерностей называют популяцией хромосом. Далее для реализации концепции отбора вводится способ сопоставления различных хромосом. Популяция обрабатывается с помощью процедур репродукции, изменчивости (мутаций), генетической композиции. Эти процедуры имитируют биологические процессы. Наиболее важные среди них: случайные мутации данных в индивидуальных хромосомах, переходы (кроссинговер) и рекомбинация генетического материала, содержащегося в индивидуальных родительских хромосомах (аналогично гетеросексуальной репродукции), и миграции генов. В ходе работы процедур на каждой стадии эволюции получаются популяции со все более совершенными индивидуумами.

Генетические алгоритмы удобны тем, что их легко распараллеливать. Например, можно разбить поколение на несколько групп и работать с каждой из них независимо, обмениваясь время от времени несколькими хромосомами. Существуют также и другие методы распараллеливания генетических алгоритмов.

Генетические алгоритмы имеют ряд недостатков. Критерий отбора хромосом и используемые процедуры являются эвристическими и далеко не гарантируют нахождения "лучшего" решения. Как и в реальной жизни, эволюцию может "заклинить" на какой-либо непродуктивной ветви. И, наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Это особенно становится заметно при решении высокоразмерных задач со сложными внутренними связями.

Примером может служить система GeneHunter фирмы Ward Systems Group. Его стоимость - около $1000.

## 4.8 Алгоритмы ограниченного перебора

Алгоритмы ограниченного перебора были предложены в середине 60-х годов М.М. Бонгардом для поиска логических закономерностей в данных. С тех пор они продемонстрировали свою эффективность при решении множества задач из самых различных областей.

Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: X = a; X < a; X a; a < X < b и др., где X - какой либо параметр, "a" и "b" - константы. Ограничением служит длина комбинации простых логических событий (у М. Бонгарда она была равна 3). На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и пр.

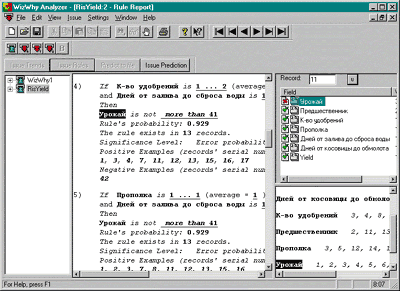
Наиболее ярким современным представителем этого подхода является система WizWhy предприятия WizSoft. Хотя автор системы Абрахам Мейдан не раскрывает специфику алгоритма, положенного в основу работы WizWhy, по результатам тщательного тестирования системы были сделаны выводы о наличии здесь ограниченного перебора (изучались результаты, зависимости времени их получения от числа анализируемых параметров и др.).

Автор WizWhy утверждает, что его система обнаруживает ***ВСЕ*** логические if-then правила в данных. На самом деле это, конечно, не так. Во-первых, максимальная длина комбинации в if-then правиле в системе WizWhy равна 6, и, во-вторых, с самого начала работы алгоритма производится эвристический поиск простых логических событий, на которых потом строится весь дальнейший анализ. Поняв эти особенности WizWhy, нетрудно было предложить простейшую тестовую задачу, которую система не смогла вообще решить.

Другой момент - система выдает решение за приемлемое время только для сравнительно небольшой размерности данных.

Тем не менее, система WizWhy является на сегодняшний день одним из лидеров на рынке продуктов Data Mining.

Это не лишено оснований. Система постоянно демонстрирует более высокие показатели при решении практических задач, чем все остальные алгоритмы. Стоимость системы около $ 4000, количество продаж - 30000.



**Рисунок 7. Система WizWhy обнаружила правила, объясняющие низкую урожайность некоторых сельскохозяйственных участков**

## 4.9 Системы для визуализации многомерных данных

В той или иной мере средства для графического отображения данных поддерживаются всеми системами Data Mining. Вместе с тем, весьма внушительную долю рынка занимают системы, специализирующиеся исключительно на этой функции. Примером здесь может служить программа DataMiner 3D словацкой фирмы Dimension5 (5-е измерение).

В подобных системах основное внимание сконцентрировано на дружелюбности пользовательского интерфейса, позволяющего ассоциировать с анализируемыми показателями различные параметры диаграммы рассеивания объектов (записей) базы данных. К таким параметрам относятся цвет, форма, ориентация относительно собственной оси, размеры и другие свойства графических элементов изображения. Кроме того, системы визуализации данных снабжены удобными средствами для масштабирования и вращения изображений. Стоимость систем визуализации может достигать нескольких сотен долларов.



**Рисунок 8. Визуализация данных системой DataMiner 3D**

## 5. Резюме

1. Рынок систем Data Mining экспоненциально развивается. В этом развитии принимают участие практически все крупнейшие корпорации (см. например http://www.kdnuggets.com). В частности, Microsoft непосредственно руководит большим сектором данного рынка (издает специальный журнал, проводит конференции, разрабатывает собственные продукты).

2. Системы Data Mining применяются по двум основным направлениям:

1) как массовый продукт для бизнес-приложений;

2) как инструменты для проведения уникальных исследований (генетика, химия, медицина и пр.). В настоящее время стоимость массового продукта от $1000 до $10000. Количество инсталляций массовых продуктов, судя по имеющимся сведениям, сегодня достигает десятков тысяч. Лидеры Data Mining связывают будущее этих систем с использованием их в качестве интеллектуальных приложений, встроенных в корпоративные хранилища данных.

3. Несмотря на обилие методов Data Mining, приоритет постепенно все более смещается в сторону логических алгоритмов поиска в данных if-then правил. С их помощью решаются задачи прогнозирования, классификации, распознавания образов, сегментации БД, извлечения из данных "скрытых" знаний, интерпретации данных, установления ассоциаций в БД и др. Результаты таких алгоритмов эффективны и легко интерпретируются.

4. Вместе с тем, главной проблемой логических методов обнаружения закономерностей является проблема перебора вариантов за приемлемое время. Известные методы либо искусственно ограничивают такой перебор (алгоритмы КОРА, WizWhy), либо строят деревья решений (алгоритмы CART, CHAID, ID3, See5, Sipina и др.), имеющих принципиальные ограничения эффективности поиска if-then правил. Другие проблемы связаны с тем, что известные методы поиска логических правил не поддерживают функцию обобщения найденных правил и функцию поиска оптимальной композиции таких правил. Удачное решение указанных проблем может составить предмет новых конкурентоспособных разработок.

## Литература

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. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? - Tandem Computers Inc., 1996.
3. Кречетов Н. Продукты для интеллектуального анализа данных. - Рынок программных средств, N14-15\_97, c.32-39.
4. Boulding K. E. General Systems Theory - The Skeleton of Science // Management Science, 2, 1956.
5. Гик Дж., Прикладная общая теория систем. - М.: Мир, 1981.
6. Киселев М., Соломатин Е. Средства добычи знаний в бизнесе и финансах. - Открытые системы, № 4, 1997, с.41-44.
7. Дюк В.А. Обработка данных на ПК в примерах. - СПб: Питер, 1997.