Введение в добычу данных (Data Mining)

Станислав Богатырёв, Александра Симонова

6 ноября 2006 г.

Содержание

Аннотация

Этот краткий доклад должен помочь осознанию сути такого сложного и многогранного явления, как Data Mining. Ставится целью довести до слушателя, что добыча данных это не просто красивая торговая марка, но и набор достаточно эффективных технологий обработки информации.

1 Коротко о главном

1.1 Что есть Data Mining...

Мы живем в веке информации. Трудно переоценить значение данных, которые мы непрерывно собираем в процессе нашей деятельности, в управлении бизнесом или производством, в банковском деле, в решении научных, инженерных и медицинских задач. Мощные компьютерные системы, хранящие и управляющие огромными базами данных, стали неотъемлемым атрибутом жизнедеятельности, как крупных корпораций, так и даже небольших компаний. Тем не менее, наличие данных само по себе еще недостаточно для улучшения показателей работы. Нужно уметь трансформировать "сырые"данные в полезную для принятия важных бизнес решений информацию. В этом и состоит основное предназначение технологий Data mining.

- Какие товары предлагать данному покупателю?

- Какова вероятность того, что данный сектор потенциальных клиентов отреагирует на рекламную кампанию?
- Можно ли выработать оптимальную стратегию игры на бирже?
- Можно ли выдать кредит данному клиенту банка?
- Какой диагноз поставить данному пациенту?
- Как прогнозировать пиковые нагрузки в телефонных или энергетических сетях?
- В чем причины брака в производственной продукции?

Ответы на эти и многие другие вопросы, возможно, содержаться в гигабайтах баз данных. Нахождение скрытых закономерностей в данных, взаимосвязей между различными переменными в базах данных, моделирование и изучение сложных систем на основе истории их поведения – вот предмет и задачи data mining. Результаты data mining – эмпирические модели, классификационные правила, выделенные кластеры и т.д. – можно затем инкорпорировать в существующие системы поддержки принятия решений и использовать их для прогноза будущих ситуаций.

Суть и цель технологии Data Mining можно выразить в нескольких словах: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и практически полезных закономерностей, а так же их проверки на новых наборах данных.

Неочевидных — это значит, что найденные закономерности не обнаруживаются стандартными статистическими методами обработки информации или даже опытными экспертами. Дело в том, что стандартные статистические методы преимущественно ориентированы лишь на обобщение информации, а не ее глубокий анализ. Эксперты же будут искать закономерности на основе своего прошлого опыта. Если закономерность не укладывается в его представление, он ее никогда не обнаружит.

Объективных — это значит, что обнаруженные закономерности будут полностью соответствовать действительности в отличие, например, от экспертного мнения, которое всегда основано на субъективном и, следовательно, ограниченном, видении ситуации.

Практически полезных – это значит, что полученные выводы имеют свое конкретное бизнес-значение, которое позволит повысить прибыльность бизнеса.

1.2 . . . и чем он не является

Следует помнить о том, что Data Mining это не набор всемогущих магических практик и он не станет вдруг находить интересные вещи в базах данных и делать из них нетривиальные выводы. Так же он не освобождает от необходимости самостоятельно разбираться в предметной области и от знания бизнеса. Data Mining лишь помогает аналитику в поиске этих самых интересных закономерностей и связей, однако не оценивает их важности и не проверяет их в условиях реального мира. Так же найденные закономерности не объясняют причин какого-либо действия или поведения. Эффективность добычи данных зависит от степени понимания этих данных человеком, от того, какие данные будут исключены, от того как данные будут закодированы (поля в таблицах, избыточность данных, etc), каким алгоритмам анализа было отдано предпочтение. Data Mining не принимает решений и не даёт готовых решений проблем. Он не даст ответ на вопрос «Как мне улучшить эффект от прямых почтовых рассылок?», но поможет характеризовать людей, которые на рассылки реагируют, или не просто реагируют, а начинают покупать активнее обычного. Но и для этих двух случаев закономерности будут найдены совсем разные.

Data Mining это не роботизированная замена бизнес аналитику или менеджеру, но новый мощный инструмент ему в помощь. Любая компания, знающая свой бизнес и своих клиентов и так уже в курсе типичных для этой области закономерностей, выявленных сотрудниками за годы работы. Что Data Mining может делать действительно хорошо, так это подтверждать такие эмпирические наблюдения и находить новые как инкрементное улучшение (или способствовать прорывам при определённой доле удачи и божественного озарения).

1.3 Определения Data Mining

Существует множество определений Data mining, но в целом они совпадают в выделении 4-х основных признаков. Вот определение, которое дал Григорий Пиатецкий-Шапиро (G. Piatetsky-Shapiro), один из ведущих мировых экспертов в области Data Mining:

«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.»

Однако существуют и другие определения.

- «Data Mining – это процесс выделения из данных неявной и неструк-

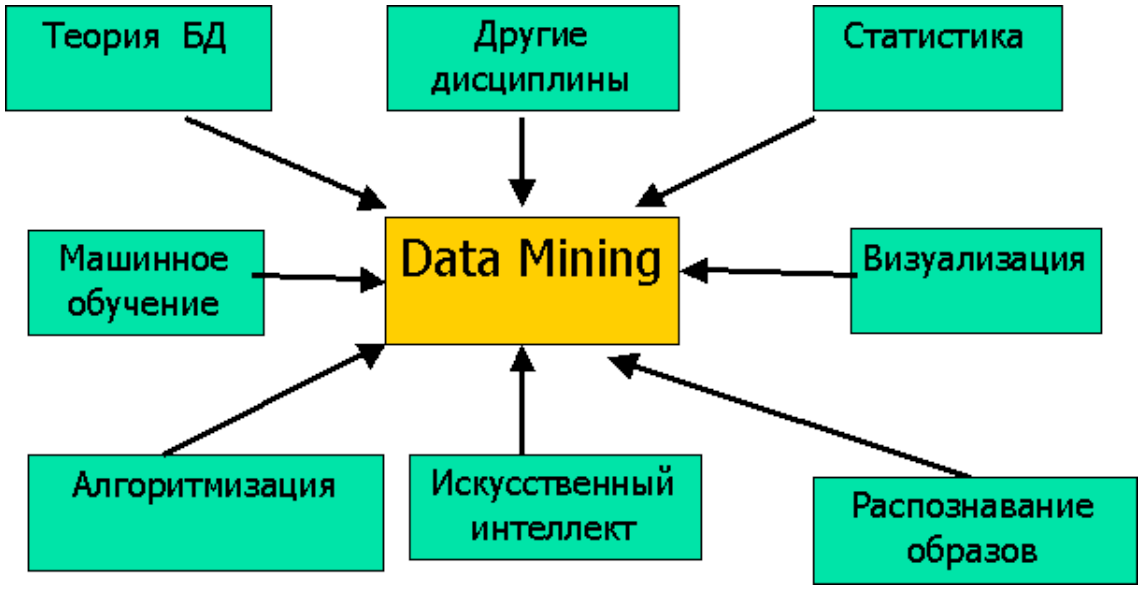


Рис. 1: Data Mining как мультидисциплинарная область

турированной информации и представления ее в виде, пригодном для использования.» (Hand)

- «Data Mining – это междисциплинарная область знаний, объединяющая технологии машинного обучения, распознавания структур (patterns), статистики, теории баз данных и визуализации с целью извлечения полезной информации из больших хранилищ данных.» (Evangelos Simoudis in Cabena)
- Data mining — это процесс выделения (selecting), исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе (SAS Institute).

1.4 Задачи Data Mining

Задачи (tasks) Data Mining иногда называют закономерностями (regularity) или техниками (techniques).

Единого мнения относительно того, какие задачи следует относить к Data Mining, нет. Большинство авторитетных источников перечисляют следующие: классификация, кластеризация, прогнозирование, ассоциация, визуализация, анализ и обнаружение отклонений, оценивание, анализ связей, подведение итогов.

1.5 Популярные примеры

Можно привести один известный пример. Data mining часто применяется в розничной торговле для выявления товаров, которые покупатели приобретают совместно, за одну покупку. Зная такие товары, специалисты выставляют их на полках рядом друг с другом и, таким образом, покупатель, купив один товар, не забудет купить и другой. Это удобно для всех – покупатели уходят довольные, продажи магазина растут. Таким образом, специалисты одного из супермаркетов крупнейшей международной сети Wal-Mart, благодаря применению data mining, обнаружили, что в пятницу вечером пиво почему-то особенно хорошо продается вместе с детскими подгузниками.

Сначала они были крайне удивлены: казалось бы, какая может быть связь между такими разными товарами? Вскоре, тем не менее, объяснение нашлось: многие мужчины, возвращаясь по вечерам с работы, по просьбе жен, покупали своим детям подгузники; а в пятницу при этом они справедливо замечали, что за тяжелую трудовую неделю заслужили свою награду – и добавляли в корзину пиво. Менеджеры Wal-Mart умело воспользовались такой находкой: поставив на полки рядом с подгузниками одни из самых дорогих марок пива, удалось добиться значительного роста его продаж.

Согласитесь, чтобы обнаружить такую зависимость при помощи статистических методов, надо было заранее предполагать, что между подгузниками и пивом может быть какая-нибудь связь. На это способны либо крайне проницательные специалисты, либо сумасшедшие. И тех, и других среди бизнес-специалистов немного, поэтому data mining является важным дополнительным инструмент повышения эффективности бизнеса.

Интересный пример использования DM привел Кирилл Резник, президент фирмы «Контекст», разработавшей пакет «ДА-система»: *«У нас была возможность поработать с крупной московской фирмой, производящей рабочую одежду. Они выставляли 50 тыс. счетов в квартал. Заказчики приезжали к ним со всей России. Вдруг они обнаружили, что примерно 30% счетов не оплачивается. А это — потраченное время на работу с клиентом, каталогами и т. д., поэтому фирма захотела узнать, какие параметры влияют на оплату счета. Было обнаружено, что очень сильно различаются клиенты из Москвы, Московской области и регионов. Клиенты из регионов приезжали под выходные, чтобы в Москве погулять, а в понедельник—вторник занимались делом. Клиенты из Москвы выставляли счета в основном в среду—четверг, а из Московской области — во вторник—среду—четверг. Мы построили*

иерархию регионов лидеров и регионов аутсайдеров. Например, Бурятия и республика Саха-Якутия всегда на протяжении четырех лет аккуратно оплачивали счета. В результате мы определили, что если клиент приходит в пятницу и выставляется счет на сумму от и до, то с вероятностью 99% счет не будет оплачен. Это не значит, что на таких людей вообще не нужно тратить время, но его можно тратить более эффективно».

1.6 История развития

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.

Понятие Data Mining, появившееся в 1978 году, приобрело высокую популярность в современной трактовке примерно с первой половины 1990-х годов. До этого времени обработка и анализ данных осуществлялся в рамках прикладной статистики, при этом в основном решались задачи обработки небольших баз данных.

Развитие технологии баз данных

- *1960-е гг.*

  В 1968 году была введена в эксплуатацию первая промышленная СУБД система IMS фирмы IBM.

- *1970-е гг.*

  В 1975 году появился первый стандарт ассоциации по языкам систем обработки данных - Conference on Data System Languages (CODASYL), определивший ряд фундаментальных понятий в теории систем баз данных, которые до сих пор являются основополагающими для сетевой модели данных. В дальнейшее развитие теории баз данных большой вклад был сделан американским математиком Э.Ф. Коддом, который является создателем реляционной модели данных.

- *1980-е гг.*

  В течение этого периода многие исследователи экспериментировали с новым подходом в направлениях структуризации баз данных и обеспечения к ним доступа. Целью этих поисков было получение реляционных прототипов для более простого моделирования данных. В результате, в 1985 году был создан язык, названный SQL.

На сегодняшний день практически все СУБД обеспечивают данный интерфейс.

- *1990-е гг.*

  Появились специфичные типы данных - "графический образ "документ "звук "карта". Типы данных для времени, интервалов времени, символьных строк с двухбайтовым представлением символов были добавлены в язык SQL. Появились технологии DataMining, хранилища данных, мультимедийные базы данных и web-базы данных. Возникновение и развитие Data Mining обусловлено различными факторами, основными среди которых являются следующие:

  – совершенствование аппаратного и программного обеспечения;
  – совершенствование технологий хранения и записи данных;
  – накопление большого количества ретроспективных данных;
  – совершенствование алгоритмов обработки информации.

2 Из жизни процесса

С ростом популярности возникла ситуация, когда многие фирмы начали «изобретать велосипед» и дублировать в своих исследованиях уже достигнутые результаты. Возникла потребность стандарте, не зависящем от конкретных программных средств, подходящем разным отраслям. В 1996 году группа аналитиков, представляющих компании DaimlerChrysler, SPPS, и NCR разработали стандарт CRISP-DM (The Cross-Industry Standard Process for Data Mining). CRISP предоставляет свободно доступный стандартный процесс для включения Data Mining в общую стратегию принятия решений организации либо её подразделения.

Сейчас действует версия 1.0 стандарта, готовится версия 2.0. Конференция, посвящённая обсуждению новой версии пройдёт [1] 26 сентября 2006 в Чикаго.

Согласно предлагаемому процессу полученная на одном этапе информация должна служить входными даными для следующего этапа обработки. Такой подход позволяет при обнаружении проблем на одной из фаз легко вернуться назад на необходимое число шагов для устранения ошибок или внесения улучшений.

[1] Или уже прошла если этот текст читается слишком поздно :-)

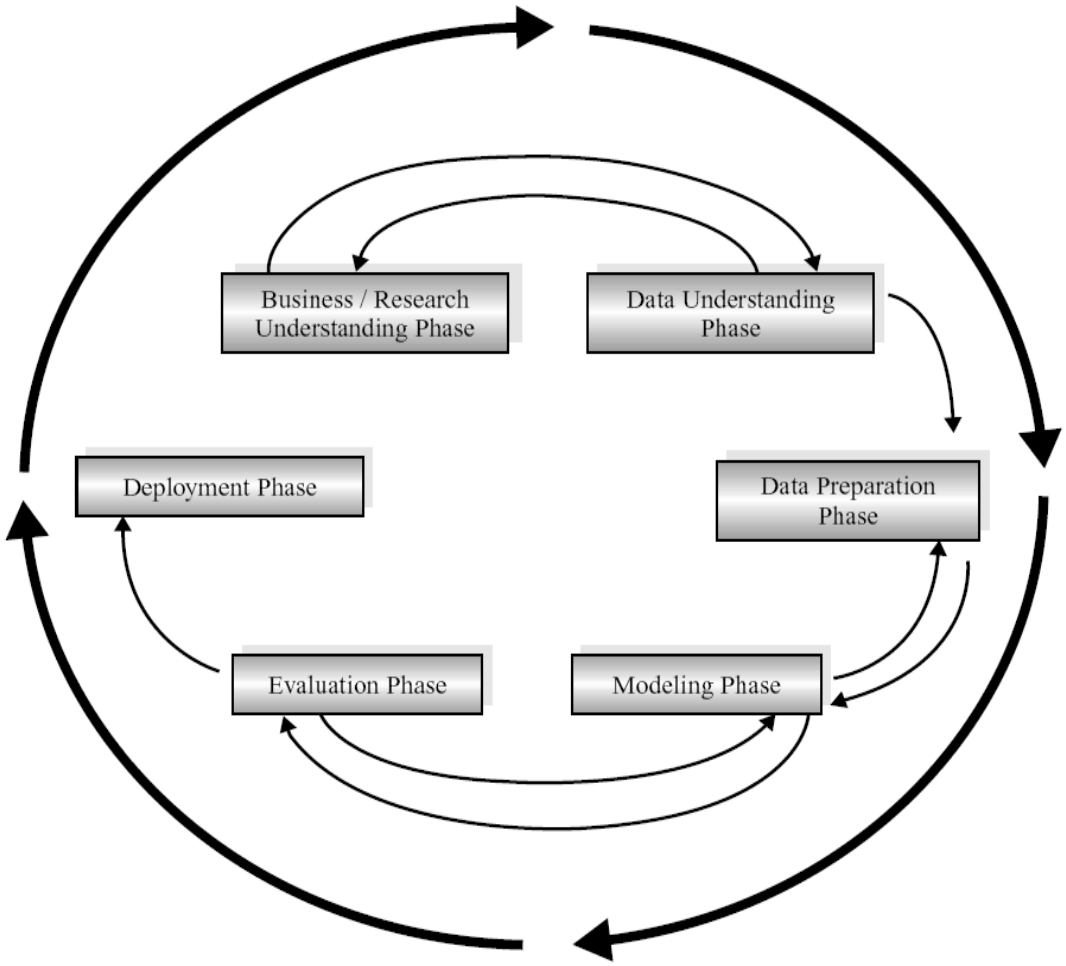


Рис. 2: Схема процесса по CRISP-DM

2.1 Фазы CRISP-DM

1. Осмысление бизнеса (Business understanding). Первая стадия, так же иногда называется фазой осмысления исследования.

   - Ясно сформулировать цели и требования проекта в понятиях бизнеса или исследования вцелом.
   - Перевести полученные цели и ограничения в формулировку проблемы добычи данных.
   - Подготовить предварительную стратегию для достижения поставленной цели.

2. Осмысление данных (Data understanding).

   - Собрать данные для обработки
   - Провести общий анализ даных что бы лучше понять их структуру, добиться начального из понимания.
   - Оценить качество имеющихся данных.
   - Возможно выбрать интересные наборы данных, в которых прослеживаются полезные нам шаблоны (patern).

3. Подготовка данных (Data preparation).

   - Подготовить из сырых данных тот набор, который будет использоваться в последующих фазах. Оди из наиболее трудоёмких этапов.
   - Выбрать переменные и случаи, подходящие для анализа.
   - При необходимости произвести трансформацию некоторых переменных.
   - Очистить сырые данные что бы окончательно подготовить их к анализу.

4. Моделирование (Modeling).

   - Выбрать и применить подходящие техники моделирования.
   - Откалибровать параметры модели для улучшения результата.
   - Подумать о том, что для решения проблемы могут быть использованы разные подходы.
   - При необходимости вернуться на предыдущую стадию и произвести переподготовку данных для лучшего соответствия новой выбранной модели.

5. Оценка результатов (Evaluation).
    - Оценить качество и эффективность моделей, выбранных на прошлом этапе.
    - Убедиться что модель действительно позволяет достигнуть целей, поставленных на первом этапе.
    - Установить что объясняются все важные грани бизнес задачи или исследования.
    - Прийти к окончательному решению об использовании результатов добычи данных.
6. Внедрение (Deployment).
    - Использовать полученную модель. Просто выработка одели не является завершением проекта.
    - В простейшем случае получить отчёт.
    - В более сложном случае провести data mining в смежных департаментах
    - Если data mining делался на заказ, то внедрение разработанной модели часто делает сам клиент.

2.2 Data Mining в действии

Что бы подробнее рассмотреть предложенный процесс рассмотрим примерный анализ данных для службы контроля качества абстрактной автомобильной компании DBL.

2.2.1 Business understanding

Руководству компании хочется уменьшить расходы, связанные с гарантийными претензиями клиентов. После разговоров с инженерами исследовательской группой были сформулированы следующие вопросы:

- Есть ли связь между гарантийными претензиями?
- Связаны ли претензии в прошлом с аналогичными случаями будущих претензий?
- Есть ли связь между типами поломок и сервисными центрами?

Планируется применить Data Mining что бы попробовать раскрыть эти, а возможно и другие, закономерности.

2.2.2 Data understanding

Фирма DBL располагает базой даных, содержащей информацию обо всех произведённых автмобилях за последние годы. В ней содержатся записи о том, где и как была произведена каждая машина, а так же гарантийные жалобы с указанием автосервиса из которого они поступили. Поломки типизированы и каждому типу присвоены идентификаторы.

При анализе баз данных исследователи пришли к выводу что информация хранится в виде непригодном для удобного использования не специалистами в этой области. Это затрудняло использование базы сотрудниками других отделов в том состоянии, в котром она была.

2.2.3 Data preparation

База данных имела структуру, оптимизированную для работы с внутренним ПО, плохо оптимизирована для SQL запросов общего назначения. Например для выяснения значения переменой «количества дней, прошедших от продажи до первого обращения в автосервис», требуется делать сложные запросы, производить обработку атрибутов даты.

Были выявлены проблемы с форматами данных. Разные алгоритмы анализа требуют данные в разном виде. Пришлось приводить всё к формату, пригодному для использования во всех планируемых к использованию алгоритмах.

2.2.4 Modeling

Так как на первом этапе была поставлена задача проследить зависимости между гарантийными рекламациями, исследовательская группа решила пользоваться следующими техниками анализа данных: сети Байеса и ассоциативные правила (association rules). Эти методики моделируют неопределённость через явное представление условных зависимостей между различными компонентами, что позволяет построить графическое представление зависимостей между объектами.

После проведения моделирования исследователи обратили внимание на то, что определённая конструкция машин удваивает вероятность возникновения проблем с электропроводкой.

Так же провели исследование получали ли какие-то сервисы больше рекламаций, связанных с определёнными типами поломок, чем другие. Ассоциативные правила показали что верность правила «Если автосервис X, значит проблема с электропроводкой» сильно менялся от сервиса к сервису. Было решено что последующие исследования должны прояснить причины таких несоответствий.

2.2.5 Evaluation

Исследовательская группа была несколько расстроена плохой поддержкой ассоциативных правил в используемом ПО, что, по их мнению, препятствовало обобщению результатов. Так же было заключено что результатов, которые могут показаться экспертной группе интересными, получено не было.

В соответствии с этим заключением использованные модели были признаны малоэффективными и не выполняющими целей, сформулированных в первой фазе. Причинами этого были указаны плохая структура базы данных, в которой детали автомобилей были категоризованы по автомастерским и заводам, по историческим или техническим причинам, что малопригодно для Data Mining. В результате была выработаны рекомендации по улучшению структуры базы данных, что позволило бы сделать её использование для получения знаний эффективнее.

2.2.6 Deployment

Исследовательская группа признала проект пилотным и решила не внедрять полученные модели. Однако после проекта они осознали полученные уроки и наметили пути интеграции новых методов с существующей IT-средой компании.

2.2.7 Выводы

Какие же выводы можно сделать из этой небольшой иллюстрации процесса добычи данных? Во-первых то, что процесс этот не прост. Практически на каждой фазе исследователи сталкивались с неожиданными проблемами и различного рода трудностями. Так же мы увидели что первая попытка внедрения data mining в компании требует от людей новых действий, что не всегда воспринимается ими с энтузиазмом. И, конечно же, если руководство компании хочет добиться результатов, то оно должно полностью поддерживать новую инициативу.

Следующий урок, извлекаемый из этого примера, это то, что на каждом этапе необходимы контроль и активное участие человека. Например, алгоритмы требуют вполне определённых форматов входной информации для своей работы, а это потребует производить дополнительную обработку данных. Так что не получится просто купить дорогое ПО и спокойно наблюдать как оно решает все проблемы как по волшебству. Без грамотного человеческого контроля слепое использование Data Mining даст неправильный ответ на не тот вопрос, руководствуясь анализом

неподходящих данных. Плохой анализ хуже вообще отсутствия анализа, так как не приведёт к принятию неверных решений, которые могу повлечь за собой серьёзные финансовые потери.

И, в конце концов, мы можем вынести для себя из этого примера, что нет никакой гарантии удачного завершения Data Minig'a :-) Хотя извлечь пользу можно, при грамотном его использовании грамотными людьми, понимающими модели, вовлечёные в процесс, требования к данным, и чётко осознающим цели проекта.

3 Методы добычи данных

Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

К методам и алгоритмам Data Mining относятся следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ; иерархические методы кластерного анализа, неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы; методы поиска ассоциативных правил, в том числе алгоритм Apriori; метод ограниченного перебора, эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и множество других методов. Классификация методов рассмотрена ниже.

3.1 Классификация методов

Все методы Data Mining подразделяются на две большие группы по принципу работы с исходными обучающими данными. В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после Data Mining либо они дистиллируются для последующего использования.

1. *Непосредственное использование данных, или сохранение данных.*

   В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях прогностического моделирования и/или анализа исключений. Проблема этой

группы методов - при их использовании могут возникнуть сложности анализа сверхбольших баз данных.

Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии.

2. *Выявление и использование формализованных закономерностей, или дистилляция шаблонов.*

   При технологии дистилляции шаблонов один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого метода Data Mining. Этот процесс выполняется на стадии свободного поиска, у первой же группы методов данная стадия в принципе отсутствует. На стадиях прогностического моделирования и анализа исключений используются результаты стадии свободного поиска, они значительно компактнее самих баз данных. Конструкции этих моделей могут быть трактуемыми аналитиком либо нетрактуемыми ("черными ящиками").

   Методы этой группы: логические методы; методы визуализации; методы кросс-табуляции; методы, основанные на уравнениях.

Методы Data Mining также можно классифицировать по задачам Data Mining.

В соответствии с такой классификацией выделяются две группы. Первая из них - это подразделение методов Data Mining на решающие задачи сегментации (т.е. задачи классификации и кластеризации) и задачи прогнозирования.

В соответствии со второй классификацией по задачам методы Data Mining могут быть направлены на получение описательных и прогнозирующих результатов.

Описательные методы служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика.

К методам, направленным на получение описательных результатов, относятся итеративные методы кластерного анализа, в том числе: алгоритм k-средних, k-медианы, иерархические методы кластерного анализа, самоорганизующиеся карты Кохонена, методы кросс-табличной визуализации, различные методы визуализации и другие.

Прогнозирующие методы используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.

К методам, направленным на получение прогнозирующих результатов, относятся такие методы: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод опорных векторов и др.

3.2 Статистические методы Data mining

В эти методы представляют собой четыре взаимосвязанных раздела:

- предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
- 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).

3.3 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 (в т.ч. алгоритмы метода группового учета аргументов);
- генетические алгоритмы (оптимизация);
- ассоциативная память (поиск аналогов, прототипов);
- нечеткая логика;
- деревья решений;
- системы обработки экспертных знаний.

4 Подробнее о некоторых методах

4.1 Деревья решений (decision trees)

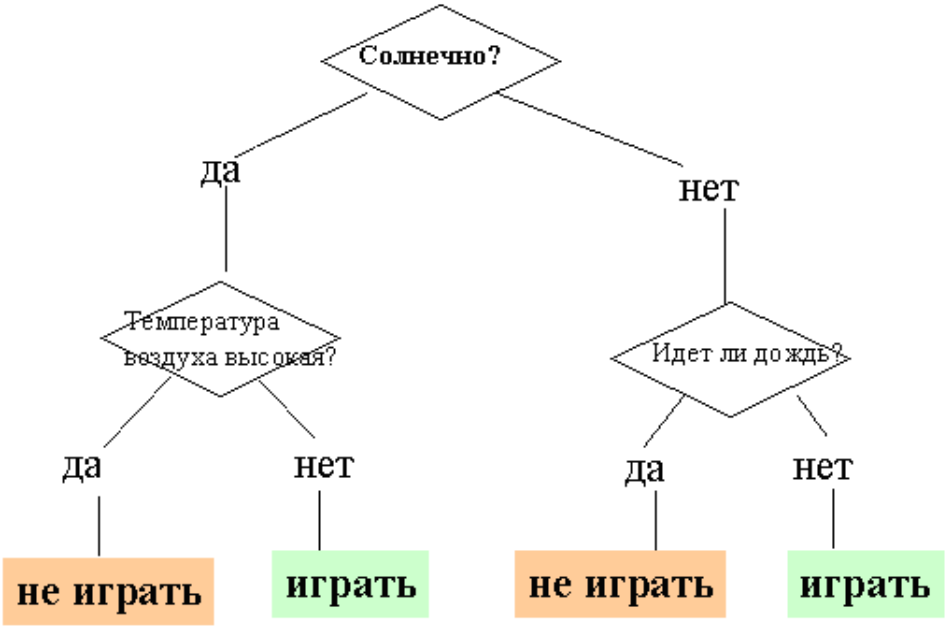
4.1.1 Определение

Этот метод Data Mining также называют деревьями решающих правил, деревьями классификации и регрессии, так как он применяется для решения задач:

1. классификации - если целевая переменная принимает дискретные значения
2. прогнозирования - если зависимая переменная принимает непрерывные значения

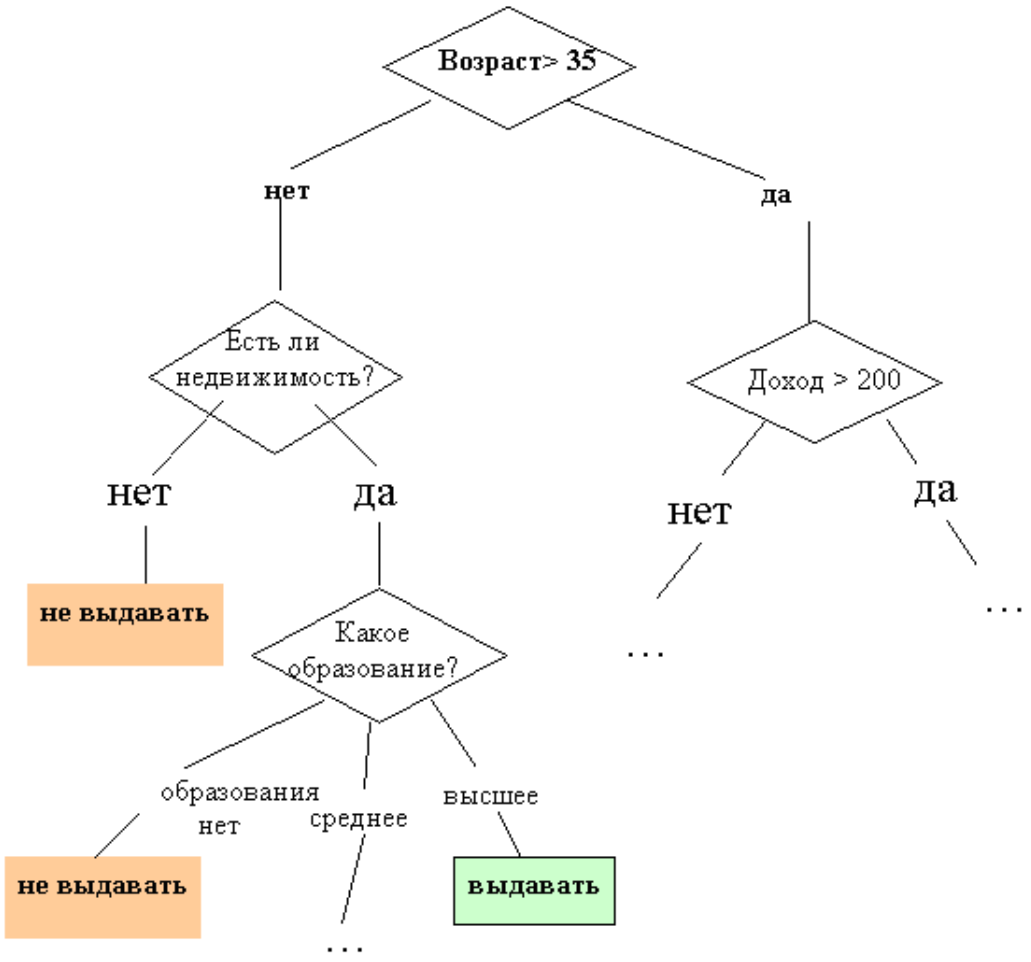
Деревья решений были предложены Ховилендом и Хантом (Hoveland, Hunt) в конце 50-х годов прошлого века. В наиболее простом виде дерево решений - это способ представления правил в иерархической, последовательной структуре.

Примеры



Пример, задача которого - ответить на вопрос: “Играть ли в гольф?” Чтобы решить задачу, т.е. принять решение, играть ли в гольф, следует отнести текущую ситуацию к одному из известных классов (в данном случае - “играть” или “не играть”). Для этого требуется ответить на ряд вопросов, которые находятся в узлах этого дерева, начиная с его корня:

1. Первый - **узел проверки** (условие)
2. При положительном ответе на вопрос осуществляется переход к левой части дерева - **левой ветви**
3. При отрицательном - к **правой части дерева**
4. Внутренний узел дерева является узлом проверки определенного условия. Далее идет следующий вопрос и т.д., пока не будет достигнут конечный узел дерева, являющийся **узлом решения**.



Это более сложный пример. База данных, на основе которой должно осуществляться прогнозирование, содержит атрибуты:

- возраст;
- наличие недвижимости;
- образование;
- среднемесячный доход;
- вернул ли клиент вовремя кредит.

Задача - на основании перечисленных выше данных (кроме последнего атрибута) определить, стоит ли выдавать кредит новому клиенту. Она решается в два этапа:

1. построение классификационной модели (дерево классификации или создается набор неких правил)

2. ее использование (набор правил для конкретного клиента - путь от корня к одной из вершин - используется для ответа на поставленный вопрос)

Каждая ветвь дерева, идущая от внутреннего узла, отмечена **предикатом расщепления**. Он может относиться лишь к одному атрибуту расщепления данного узла. Объединенная информация об атрибутах расщепления и предикатах расщепления в узле называется критерием расщепления. Качество построенного дерева решения зависит от правильного выбора **критерия расщепления**. Качество построенного дерева решения зависит от правильного выбора критерия расщепления.

4.1.2 Преимущества деревьев решений

1. Интуитивность деревьев решений. Это свойство деревьев решений не только важно при отнесении к определенному классу нового объекта, но и полезно при интерпретации модели классификации в целом.

2. Деревья решений дают возможность извлекать правила из базы данных на естественном языке.

3. Деревья решений позволяют создавать классификационные модели в областях, где достаточно сложно формализовать знания.

4. Алгоритм конструирования дерева решений не требует от пользователя выбора входных атрибутов (независимых переменных). На вход алгоритма можно подавать все существующие атрибуты, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева.

5. Точность моделей оказывается достаточно высокой.

6. Разработан ряд масштабируемых алгоритмов, которые могут быть использованы для построения деревьев решения на сверхбольших базах данных. Примеры: SLIQ, SPRINT.

7. Быстрый процесс обучения. На построение классификационных моделей при помощи алгоритмов конструирования деревьев решений требуется значительно меньше времени, чем, например, на обучение нейронных сетей.

8. Большинство алгоритмов конструирования деревьев решений имеют возможность специальной обработки пропущенных значений.

9. Деревья решений работают и с числовыми, и с категориальными типами данных.

10. Деревья решений способны решать такие задачи Data Mining, в которых отсутствует априорная информация о виде зависимости между исследуемыми данными.

4.1.3 Конструирование дерева решений

Алгоритмы конструирования деревьев решений состоят из 2 этапов:

1. "построение" или "создание" дерева (tree building) - решаются вопросы выбора критерия расщепления и остановки обучения (если это предусмотрено алгоритмом)

2. "сокращение" дерева (tree pruning) - решается вопрос отсечения некоторых его ветвей

Критерий расщепления Процесс создания дерева происходит сверху вниз, т.е. является нисходящим. Каждый узел проверки должен быть помечен определенным атрибутом. Существует **правило выбора атрибута**: он должен разбивать исходное множество данных таким образом, чтобы объекты подмножеств, получаемых в результате этого разбиения, являлись представителями одного класса или же были максимально приближены к такому разбиению.

Существуют различные критерии расщепления. Наиболее известные:

- мера энтропии (мера информативности подпространств атрибутов, которая основывается на энтропийном подходе)

- индекс Gini (атрибут выбирается на основании расстояний между распределениями классов) - $gini(T) = 1 - \sum_{j=1}^{n} p_j^2$ где T - текущий узел, p_j - вероятность класса j в узле T, n - количество классов.

Оптимальный размер дерева Какой размер дерева может считаться оптимальным? Другими словами, дерево должно использовать информацию, улучшающую качество модели, и игнорировать ту информацию, которая ее не улучшает.

Существуют две возможные стратегии:

1. наращивание дерева до определенного размера в соответствии с параметрами, заданными пользователем

2. использование набора процедур, определяющих "подходящий размер" дерева

Процедуры, использующиеся для предотвращения создания чрезмерно больших деревьев, включают:

1. сокращение дерева путем отсечения ветвей

2. использование правил остановки обучения

Остановка построения дерева Правило остановки определяет, является ли рассматриваемый узел внутренним узлом и будет разбиваться дальше, или он является конечным узлом (узлом решением).

Остановка - такой момент в процессе построения дерева, когда следует прекратить дальнейшие ветвления.

Варианты правил остановки следующие:

1. "ранняя остановка" (prepruning) - определяет целесообразность разбиения узла. Преимущество - уменьшение времени на обучение модели. Недостаток - возникает риск снижения точности классификации.

2. ограничение глубины дерева - построение заканчивается, если достигнута заданная глубина.

3. задание минимального количества примеров, которые будут содержаться в конечных узлах дерева - ветвления продолжаются до того момента, пока все конечные узлы дерева не будут чистыми или будут содержать не более чем заданное число объектов.

Сокращение дерева Качество классификационной модели, построенной при помощи дерева решений, характеризуется двумя основными признаками:

1. **Точность распознавания** - отношение объектов, правильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

2. **Ошибка** - отношение объектов, неправильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

Отсечение ветвей или замену некоторых ветвей поддеревом следует проводить там, где эта процедура не приводит к возрастанию ошибки.

4.1.4 Алгоритмы

На сегодняшний день существует большое число алгоритмов, реализующих деревья решений: CART, C4.5, CHAID, CN2, NewId, ITrule и другие.

Алгоритм CART Алгоритм CART (Classification and Regression Tree) решает задачи классификации и регрессии. Он разработан в 1974-1984 годах четырьмя профессорами статистики - Leo Breiman (Berkeley), Jerry Friedman (Stanford), Charles Stone (Berkeley) и Richard Olshen (Stanford).

Особенности алгоритма CART:

- атрибуты набора данных могут иметь как дискретное, так и числовое значение
- алгоритм предназначен для построения бинарного дерева решений
- функция оценки качества разбиения
- механизм отсечения дерева
- алгоритм обработки пропущенных значений
- построение деревьев регрессии
- используемая функция оценки качества разбиения - индекс Gini
- правила разбиения -
  в каждом узле разбиение может идти только по одному атрибуту. Если атрибут является числовым, то во внутреннем узле формируется правило вида $x_i <= c$, Значение c - среднее арифметическое двух соседних упорядоченных значений переменной x_i обучающего набора данных. Если же атрибут относится к категориальному типу, то во внутреннем узле формируется правило $x_i V(x_i)$, где $V(x_i)$ - некоторое непустое подмножество множества значений переменной x_i в обучающем наборе данных.
- механизм отсечения - это некий компромисс между получением дерева “подходящего размера” и получением наиболее точной оценки классификации. Метод заключается в получении последовательности уменьшающихся деревьев, но деревья рассматриваются не все, а только “лучшие представители”.

- перекрестная проверка (V-fold cross-validation) - путь выбора окончательного дерева, при условии, что набор данных имеет небольшой объем или же записи набора данных настолько специфические, что разделить набор на обучающую и тестовую выборку не представляется возможным.

Алгоритм C4.5 Особенности алгоритма C4.5:

1. построение дерева решений с неограниченным количеством ветвей у узла
2. работа только с дискретным зависимым атрибутом
3. медленная работа на сверхбольших и зашумленных наборах данных
4. робастость

Для работы алгоритма C4.5 необходимо соблюдение следующих требований:

- Каждая запись набора данных должна быть ассоциирована с одним из предопределенных классов, т.е. один из атрибутов набора данных должен являться меткой класса.
- Классы должны быть дискретными. Каждый пример должен однозначно относиться к одному из классов.
- Количество классов должно быть значительно меньше количества записей в исследуемом наборе данных.

Алгоритмы построения деревьев решений различаются следующими характеристиками:

- вид расщепления - бинарное (binary), множественное (multi-way)
- критерии расщепления - энтропия, Gini, другие
- возможность обработки пропущенных значений
- процедура сокращения ветвей или отсечения
- возможности извлечения правил из деревьев.

4.2 Метод "ближайшего соседа" или системы рассуждений на основе аналогичных случаев

4.2.1 Определение

Метод "ближайшего соседа" ("nearest neighbour") основывается на хранении данных в памяти для сравнения с новыми элементами - при появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется. Также используется подход "k-ближайший соседей" ("k-nearest neighbour") - выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей". Иногда хранится только множество "типичных" случаев. В таком случае используемый метод называют рассуждением по аналогии (Case Based Reasoning, CBR).

Подход, основанный на прецедентах, условно можно поделить на следующие этапы:

- сбор подробной информации о поставленной задаче;
- сопоставление этой информации с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев;
- выбор прецедента, наиболее близкого к текущей проблеме, из базы прецедентов;
- адаптация выбранного решения к текущей проблеме, если это необходимо;
- проверка корректности каждого вновь полученного решения;
- занесение детальной информации о новом прецеденте в базу прецедентов.

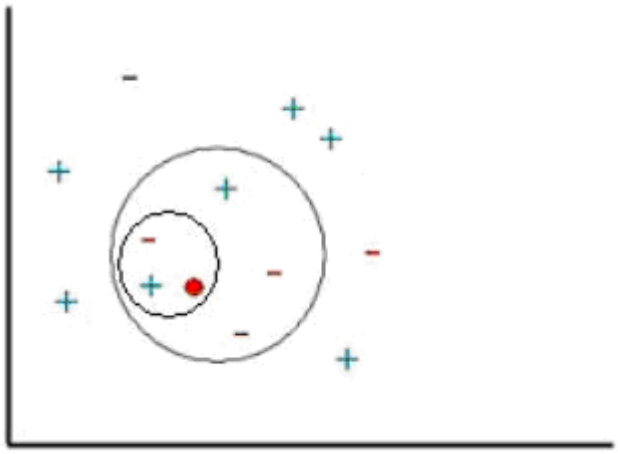
4.2.2 Преимущества метода

- Простота использования полученных результатов.
- Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
- Целью поиска является не гарантированно верное решение, а лучшее из возможных.

4.2.3 Недостатки метода

- Не создается моделей или правил, обобщающих предыдущий опыт, - в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы.
- Сложность выбора меры “близости” (метрики) - от этой меры главным образом зависит объем множества записей, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза.
- Высокая зависимость результатов классификации от выбранной метрики.
- Необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость.
- Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

4.2.4 Решение задачи классификации новых объектов



Обозначения:

1. примеры (известные экземпляры) отмечены знаком “+” или “-”, определяющим принадлежность к соответствующему классу (“+” или “-”)

2. новый объект, который требуется классифицировать, обозначен красным кружочком

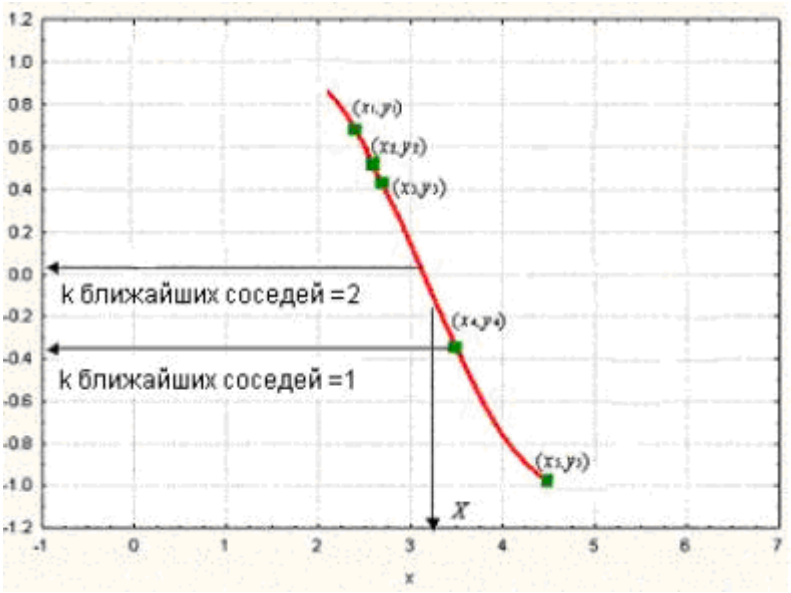
Цель - классификация отклика точек запроса с использованием специально выбранного числа их ближайших соседей. Рассмотрим различные значения k:

1. $k = 1$ - отклик точки запроса будет классифицирован как знак плюс, так как ближайшая соседняя точка имеет знак плюс

2. $k = 2$ - метод k-ближайших соседей не сможет классифицировать отклик точки запроса, поскольку вторая ближайшая точка имеет знак минус и оба знака равноценны (т.е. победа с одинаковым количеством голосов)

3. $k = 5$ - будет определена целая окрестность точки запроса (на графике ее граница отмечена красной(серой) окружностью). Так как в области содержится 2 точки со знаком “+” и 3 точки со знаком “-” , алгоритм k-ближайших соседей присвоит знак “-” отклику точки запроса.

4.2.5 Решение задачи прогнозирования



Обозначения:

1. набор точек (зеленые прямоугольники) получен по связи между
2. независимой переменной x и зависимой переменной y (кривая красного цвета)
3. задан набор зеленых объектов (т.е. набор примеров);
4. мы используем метод k-ближайших соседей для предсказания выхода точки запроса X по данному набору примеров (зеленые прямоугольники).

Рассмотрим различные значения k:

1. $k = 1$ - ищем набор примеров (зеленые прямоугольники) и выделяем из их числа ближайший к точке запроса X. Для нашего случая ближайший пример - точка (x_4, y_4). Выход x_4 (т.е. y_4), таким образом, принимается в качестве результата предсказания выхода X (т.е. Y).
2. $k = 2$ - выделяем уже две ближайшие к X точки. На нашем графике это точки y_3 и y_4 соответственно. Вычислив среднее их выходов, записываем решение для Y в виде $Y = \frac{(y_3+y_4)}{2}$.

Предсказание в задаче прогнозирования получается усреднением выходов k-ближайших соседей, а решение задачи классификации основано на принципе “по большинству голосов”.

Критический момент - выбор параметра k:

1. Если выбрано слишком маленькое значение параметра k, возникает вероятность большого разброса значений прогноза.

2. Если выбранное значение слишком велико, это может привести к сильной смещенности модели.

4.2.6 Оценка параметра k

1. проведение кросс-проверки (Bishop, 1995). Кросс-проверка - известный метод получения оценок неизвестных параметров модели. Основная идея метода - разделение выборки данных на v “складок” (V “складки” здесь суть случайным образом выделенные изолированные подвыборки). По k строится модель k-ближайших соседей для получения предсказаний на v-м сегменте. Далее процесс последовательно повторяется для всех возможных вариантов выбора v. Вышеописанные действия повторяются для различных k, и значение, соответствующее наименьшей ошибке (или наибольшей классификационной точности), принимается как оптимальное. По фиксированному значению (остальные сегменты при этом используются как примеры) и оценивается ошибка классификации. Для регрессионных задач наиболее часто в качестве оценки ошибки выступает сумма квадратов, а для классификационных задач удобней рассматривать точность (процент корректно классифицированных наблюдений). По исчерпании v “складок” (циклов), вычисленные ошибки усредняются и используются в качестве меры устойчивости модели (т.е. меры качества предсказания в точках запроса).

2. самостоятельное задание значения k.

4.2.7 Примеры использования

Примером реального использования - программное обеспечение центра технической поддержки компании Dell, разработанное компанией Inference. Эта система помогает сотрудникам центра отвечать на большее число запросов, сразу предлагая ответы на распространенные вопросы и позволяя обращаться к базе во время разговора по телефону с пользователем. Сотрудники центра технической поддержки, благодаря

реализации этого метода, могут отвечать одновременно на значительное число звонков. Программное обеспечение CBR сейчас развернуто в сети Intranet компании Dell.

4.2.8 Примеры реализации

Наиболее известные:

1. CBR Express и Case Point (Inference Corp.)
2. Apriori (Answer Systems)
3. DP Umbrella (VYCOR Corp.)
4. KATE tools (Acknosoft, Франция)
5. Pattern Recognition Workbench (Unica, США)
6. Statistica.

5 Подходя к концу

5.1 Проблемы Data Mining

Прежде чем использовать технологию Data Mining, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, с ней связанные, а также понять, чего эта технология не может.

- *Data Mining не может заменить аналитика!*

  Технология не может дать ответы на те вопросы, которые не были заданы. Она не может заменить аналитика, а всего лишь дает ему мощный инструмент для облегчения и улучшения его работы.

- *Сложность разработки и эксплуатации приложения Data Mining*

  Поскольку данная технология является мультидисциплинарной областью, для разработки приложения, включающего Data Mining, необходимо задействовать специалистов из разных областей, а также обеспечить их качественное взаимодействие.

- *Квалификация пользователя*

  Различные инструменты Data Mining имеют различную степень "дружелюбности"интерфейса и требуют определенной квалификации пользователя. Поэтому программное обеспечение должно соответствовать уровню подготовки пользователя. Использование Data

Mining должно быть неразрывно связано с повышением квалификации пользователя. Однако специалистов по Data Mining, которые бы хорошо разбирались в бизнесе, пока еще мало.

- *Извлечение полезных сведений невозможно без хорошего понимания сути данных*

  Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов, которые обнаружены. Поэтому работа с такими средствами требует тесного сотрудничества между экспертом в предметной области и специалистом по инструментам Data Mining. Построенные модели должны быть грамотно интегрированы в бизнес-процессы для возможности оценки и обновления моделей. В последнее время системы Data Mining поставляются как часть технологии хранилищ данных.

- *Сложность подготовки данных*

  Успешный анализ требует качественной предобработки данных. По утверждению аналитиков и пользователей баз данных, процесс предобработки может занять до 80% процентов всего Data Mining-процесса.

  Таким образом, чтобы технология работала на себя, потребуется много усилий и времени, которые уходят на предварительный анализ данных, выбор модели и ее корректировку.

- *Большой процент ложных, недостоверных или бессмысленных результатов*

  С помощью Data Mining можно отыскивать действительно очень ценную информацию, которая вскоре даст большие дивиденды в виде финансовой и конкурентной выгоды. Однако Data Mining достаточно часто делает множество ложных и не имеющих смысла открытий. Многие специалисты утверждают, что Data Mining-средства могут выдавать огромное количество статистически недостоверных результатов. Чтобы этого избежать, необходима проверка адекватности полученных моделей на тестовых данных.

- *Высокая стоимость*

  Качественная Data Mining-программа может стоить достаточно дорого для компании. Вариантом служит приобретение уже готового решения с предварительной проверкой его использования, например на демо-версии с небольшой выборкой данных.

- *Наличие достаточного количества репрезентативных данных*

  Средства Data Mining, в отличие от статистических, теоретически не требуют наличия строго определенного количества ретроспективных данных. Эта особенность может стать причиной обнаружения недостоверных, ложных моделей и, как результат, принятия на их основе неверных решений. Необходимо осуществлять контроль статистической значимости обнаруженных знаний.

5.2 Перспективы технологии Data Mining

Потенциал Data Mining дает "зеленый свет"для расширения границ применения технологии. Относительно перспектив Data Mining возможны следующие направления развития:

- выделение типов предметных областей с соответствующими им эвристиками, формализация которых облегчит решение соответствующих задач Data Mining, относящихся к этим областям;
- создание формальных языков и логических средств, с помощью которых будет формализованы рассуждения и автоматизация которых станет инструментом решения задач Data Mining в конкретных предметных областях;
- создание методов Data Mining, способных не только извлекать из данных закономерности, но и формировать некие теории, опирающиеся на эмпирические данные;
- преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области.

Если рассматривать будущее Data Mining в краткосрочной перспективе, то очевидно, что развитие этой технологии наиболее направлено к областям, связанным с бизнесом.

В краткосрочной перспективе продукты Data Mining могут стать такими же обычными и необходимыми, как электронная почта, и, например, использоваться пользователями для поиска самых низких цен на определенный товар или наиболее дешевых билетов.

В долгосрочной перспективе будущее Data Mining является действительно захватывающим - это может быть поиск интеллектуальными агентами как новых видов лечения различных заболеваний, так и нового понимания природы вселенной.

Однако Data Mining таит в себе и потенциальную опасность - ведь все большее количество информации становится доступным через всемирную сеть, в том числе и сведения частного характера, и все больше знаний возможно добыть из нее:

Не так давно крупнейший онлайновый магазин "Amazon"оказался в центре скандала по поводу полученного им патента "Методы и системы помощи пользователям при покупке товаров который представляет собой не что иное как очередной продукт Data Mining, предназначенный для сбора персональных данных о посетителях магазина. Новая методика позволяет прогнозировать будущие запросы на основании фактов покупок, а также делать выводы об их назначении. Цель данной методики - то, о чем говорилось выше - получение как можно большего количества информации о клиентах, в том числе и частного характера (пол, возраст, предпочтения и т.д.). Таким образом, собираются данные о частной жизни покупателей магазина, а также членах их семей, включая детей. Последнее запрещено законодательством многих стран - сбор информации о несовершеннолетних возможен там только с разрешения родителей.

Исследования отмечают, что существуют как успешные решения, использующие Data Mining, так и неудачный опыт применения этой технологии. Области, где применения технологии Data Mining, скорее всего, будут успешными, имеют такие особенности:

- требуют решений, основанных на знаниях;
- имеют изменяющуюся окружающую среду;
- имеют доступные, достаточные и значимые данные;
- обеспечивают высокие дивиденды от правильных решений.

Список литературы

[1] «Wikipedia about Data Mining», http://en.wikipedia.org/wiki/Data_mining

[2] «Data Mining Tutorials», http://www.eruditionhome.com/datamining/tut.html

[3] «INTUIT.ru: Учебный курс - Data Mining», http://www.intuit.ru/department/database/datamining/

[4] «Data Mining - подготовка исходных данных», http://www.basegroup.ru/tasks/datamining_prepare.htm

[5] *Two Crows Corp*, Introduction to Data Mining and Knowledge Discovery: Third Edition, 1999

[6] *Larose, Daniel T.*, Discovering knowledge in data. An introduction to Data Mining, 2005