Лекция 8, 9. ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Часть 1. Применение технологий ИИ для интеллектуальной обработки информации. Технология диалоговой аналитической обработки (OLAP).

Часть 2. Обзор и классификация задач Data Mining.

Состав и архитектуры современной системы интеллектуальной обработки информации. Построение систем бизнес-аналитики, инструментарий – решения от компании Microsoft.

Применение технологий ИИ для интеллектуальной обработки информации

Интеллектуальная обработка информации предусматривает преобразование необработанных данных в ценную и полезную информацию. Процесс такой «продвинутой аналитики» (англ. Advanced Analytics) позволяет искать и идентифицировать тенденции, модели поведения и паттерны в больших наборах данных с помощью широкого спектра технологий. Основными используемыми технологиями здесь выступают: методы искусственного интеллекта и машинное обучение в совокупности с технологиями хранения и обработки данных в СУБД. И, в то же время, одних этих методов и технологий не достаточно. Data Mining – это мультидисциплинарная область, которая базируется как на множестве направлений информатики, так и множестве других наук (лингвистика, (рисунок 8.1).

Рисунок 8.1. Data Mining как мультидисциплинарная область

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

небольших, по настоящим меркам, баз данных. Возникновение и развитие Data Mining обусловлено различными факторами, и основными из них являются:

– накопление большого количества ретроспективных данных и возросшие потребности бизнеса в их практическом использовании;

– совершенствование технологий хранения и манипулирования данными;

– совершенствование алгоритмов обработки информации и постоянное совершенствование инструментальных программных средств;

– совершенствование аппаратного обеспечения для поддержки обработки больших объёмов данных.

Основная цель процесса интеллектуального анализа данных (Discovery-Driven Data mining или Data Mining) заключается в обнаружении и извлечении полезной информации путём просеивания массы исходных «сырых» данных. Сырыми данными (raw data) называются необработанные данные, сокращение объема которых не производилось за счет замены некоторых отдельных частей этих данных оценочными величинами, производными от них (например, средними значениями). Будь то большие массивы текстов или наборы фотографий для распознавания, технологии интеллектуального анализа данных позволяют выявлять актуальную информацию для использования в задачах бизнеса, в социальной сфере, в медицине, военной отрасли, в образовании и т.д. Программные системы интеллектуального анализа данных дают возможность превращать разрозненные сырые данные в целостную и понятную структурированную информацию, которая уже имеет прямое назначение.

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). Компании могут использовать программное обеспечение извлечения данных для формирования пула потенциальных клиентов, сбора релевантной информации с веб-страниц конкурирующих компаний, выявления тенденций из коллекций документов и анализа неструктурированной текстовой информации. Системы интеллектуального анализа и извлечения данных могут помочь предприятиям в цифровизации бизнеса, а уже перешедшим на цифровое взаимодействие – заставить работать те неструктурированные данные, которые в настоящее время не используются. 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

Таблица 8.1 Примеры применения интеллектуального анализа данных в различных предметных областях

Задачи	Комментарии	Закономерности
Розничная торговля		
Анализ покупательской корзины (сходства)	Выявление товаров, которые стремятся покупать вместе. Применение - реклама, стратегия, создания запасов и размещения товаров	Ассоциации
Исследование временных шаблонов активности покупателей	Применение – создание товарных запасов	Анализ временны́х последовательностей
Банковское дело		
Выявление мошенничества с кредитными карточками	Выявление стереотипов поведения мошенников в результате анализа исторических данных	Классификация
Сегментация клиентов	Выявление ориентированности различных трупп клиентов на различные виды услуг	Кластеризация
Телекоммуникации		
Анализ записей подробных характеристиках вызовов	Выявление стереотипов пользования услугами и разработка привлекательных наборов цен и услуг	Кластеризация
Выявление лояльности клиентов	Описание характеристик клиентов, склонных к неоднократному пользованию услугами компании	Классификация
Страхование		
Выявление мошенничества	Выявление стереотипов поведения мошенников	Классификация
Анализ риска	Пересмотр, политики предоставления скидок в результате анализа факторов, связанных с оплаченными заявлениями	Классификация

Пожалуй, наиболее остро и вместе с тем четко задача обнаружения закономерностей в экспериментальных данных стоит в молекулярной генетике и генной инженерии. В последнее время в данной области возник особый

интерес к применению методов Data Mining. Известно несколько крупных фирм, специализирующихся на применении этих методов для решения задачи расшифровки генома человека и растений. 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.

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

Основные этапы интеллектуального анализа данных

Прежде чем использовать технологию Data Mining, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, с ней связанные, а также понять, что эта технология не может. Например, Data Mining не может заменить аналитика, она всего лишь дает ему мощный инструмент для облегчения и улучшения его работы.

В процессе интеллектуального анализа данных можно выделить следующие основные этапы:

Определение проблемы. Этот шаг включает анализ бизнес-требований, определение области проблемы, метрик, по которым будет выполняться оценка модели, а также определение задач для проекта интеллектуального анализа данных. Эти задачи можно сформулировать в виде следующих вопросов:

Что необходимо найти? Какие типы связей необходимо найти? Отражает ли решаемая задача бизнес-правила или бизнес-процессы? Надо ли делать прогнозы на основании модели интеллектуального анализа данных или просто найти содержательные закономерности и взаимосвязи? Какой результат или атрибут необходимо спрогнозировать? Какие виды данных нужно иметь и какого рода информация находится в каждом столбце? Если существует несколько таблиц, как они связаны? Нужно ли выполнять очистку, статистическую обработку или обработку, чтобы данные стали применимыми? Каким образом распределяются данные? Являются ли данные сезонными? Дают ли данные точное представление бизнес-процессов?

Выборка, объединение и очистка данных. Существенная часть работы над Data Mining состоит в подготовке и интеграции данных еще до того, как запускаются сами инструменты интеллектуального анализа. Обычно не все имеющиеся в распоряжении данные необходимы для использования в Data Mining. Выборка данных – это этап, в котором из большой базы данных выбираются и извлекаются только полезные данные. Данные из разных

источников, комбинируются и интегрируются. Источниками могут быть базы данных, текстовые файлы, электронные таблицы, документы, многомерные массивы данных, интернет и так далее.

Хранимые данные не всегда очищаются и структурируются. Часто они зашумлены, имеют пропуски, могут содержать ошибки, выбросы и аномалии. Чтобы удостовериться, что результат Data Mining точный, сначала необходимо очистить данные. Некоторые методы очистки включают заполнение недостающих значений, автоматический и ручной контроль и т.д.

Просмотр и изучение данных. Для принятия правильных решений при создании моделей интеллектуального анализа данных необходимо понимать данные. Методы исследования данных включают в себя расчет минимальных и максимальных значений, вычисление средневероятного и стандартного отклонения и изучение распределения данных. Например, по максимальному, минимальному и среднему значениям можно заключить, что выборка данных не является репрезентативной для решения поставленной задачи, и поэтому необходимо получить более сбалансированные данные или изменить предположения, лежащие в основе ожидаемых результатов. Стандартное отклонение и другие характеристики распределения могут сообщить полезные сведения о стабильности и точности результатов. Большая величина стандартного отклонения может свидетельствовать о том, что добавление новых данных поможет усовершенствовать модель. Данные, которые сильно отклоняются от стандартного распределения, могут оказаться искаженными или представлять точную картину реальной проблемы, которая делает сложным подбор соответствующей модели для данных. Изучение данных в свете собственных представлений о бизнес-проблеме может привести к выводу о наличии ошибок в наборе данных, и затем можно выработать стратегию для устранения проблем или получить более глубокое представление о моделях поведения, характерных для бизнеса.

Преобразование данных. После выбора данных они преобразуются в подходящие для последующей обработки формы. Этот процесс включает в себя нормирование, агрегирование, обобщение и т.д.

Интеллектуальный анализ данных. Здесь наступает самая важная часть Data Mining – использование интеллектуальных методов для поиска закономерностей в них. Необходимы тщательный выбор модели и настройка её параметров как до начала использования, так и в процессе создания (многочисленных экспериментов). Для формирования модели используются входные данные и параметры, управляющие алгоритмом обработки данных. Обработку модели часто называют *обучением*. Обучение обозначает процесс применения некоторого алгоритма к данным в структуре с целью выявления закономерности. Различные задачи реализуются различными моделями, часто совершенно отличную от других. После прохождения данных через модель объект модели интеллектуального анализа данных будет содержать сводные данные и закономерности, которые можно запрашивать и использовать для прогнозирования.

Исследование и оценка модели. Перед развертыванием модели в рабочей среде необходимо проверить эффективность её работы. Кроме того, во время построения модели обычно создается несколько моделей с различной конфигурацией, а затем проверяются все модели, чтобы определить, какая из них обеспечивает лучшие результаты для поставленной задачи и имеющихся данных. Как правило ещё до создания модели производится разделение данных на данные для обучения и проверочный (тестовый) набор, чтобы можно было точно оценить производительность всех моделей, основанных на одних и тех же данных. Набор данных для обучения используется в ходе построения модели, а набор проверочных данных – для проверки точности модели путем создания прогнозирующих запросов.

Тенденции и закономерности, обнаруживаемые алгоритмами, необходимо исследовать различными способами. В зависимости от типа модели исследуются различные её аспекты:

– использовать различные мер статистической достоверности с целью выявления проблем в данных или в модели;

– осуществить проверку на тестовых наборах для проверки точности прогнозов;

– обратиться к специалистам с просьбой изучить результаты модели интеллектуального анализа данных и определить, имеют ли выявленные практическую ценность.

Методы оценки используются по мере создания, проверки и уточнения модели используются многократно в зависимости от выявленных проблем. Нет исчерпывающего правила, позволяющего однозначно судить о том, какие модели можно считать достаточно хорошими и достаточно ли имеющихся данных.

Можно выделить следующие категории для оценки модели: точность, надежность и информативность. Точность – это мера того, насколько выходные данные модели соответствуют тестовым данным. Имеется несколько мер точности, но все они зависят от используемых данных. Надежность соответствует поведению модели интеллектуального анализа данных на различных наборах данных. Модель интеллектуального анализа данных считается надежной, если она формирует один и тот же тип прогнозов или находит одни и те же общие типы закономерностей, вне зависимости от предоставляемый проверочных данных. Информативность объединяет в себе несколько метрик, позволяющих понять, насколько полезна информация, получаемая из модели.

На этом этапе полученные результаты представляются в репрезентативном виде с применением различных методов визуализации.

Применение наиболее эффективных моделей. На готовой модели можно выполнять множество задач, соответствующих потребностям пользователя.

Интеллектуальный анализ данных – это длительный и сложный процесс, он требует напряженной продуктивной работы квалифицированных

специалистов. Исследователи и специалисты по интеллектуальному анализу данных могут воспользоваться мощными инструментами добычи данных, однако такая работа требует я тесного сотрудничества со специалистами (экспертами) предметной области – как на этапах подготовки данных, выбора модели, так и на этапах оценки модели и интерпретации результатов.

Построенные модели должны быть грамотно интегрированы в бизнес-процессы для возможности оценки и последующего обновления.

Технология диалоговой аналитической обработки (OLAP)

Online Analytical Processing (OLAP) – мощная технология аналитической обработки структурированных данных в реальном времени. Соответственно OLAP-системы – это информационные системы, предоставляющие практически безграничные возможности по составлению отчетов, выполнению сложных аналитических расчетов, построению прогнозов и сценариев, разработке множества вариантов планов. В основе работы OLAP системы лежит обработка многомерных массивов данных. Многомерные массивы устроены так, что каждый элемент массива имеет множество связей с другими элементами. Пользователь OLAP-системы получает необходимые данные из разных источников в структурированном виде в соответствии со своим запросом.

Полноценные OLAP системы появились в начале 1990-х гг. как результат развития информационных систем поддержки принятия решений. Эти системы предназначены для преобразования различных, часто разрозненных, данных, в полезную информацию. OLAP-системы могут организовать данные в соответствии с некоторым набором критериев. При этом не обязательно, чтобы критерии имели четкие характеристики. Свое применение OLAP системы нашли во многих вопросах стратегического управления организацией: управление эффективностью бизнеса, стратегическое планирование, бюджетирование, прогнозирование развития, подготовка финансовой отчетности, анализ работы, имитационное моделирование внешней и внутренней среды организации, хранение данных и отчетности.

OLAP системы включают следующие ключевые компоненты:

– Базу данных – источник, из которого берется информационный материал для обработки. Тип БД определяется разновидностью OLAP системы и порядком выполнения действий OLAP-сервера. Чаще всего пользуются реляционными и многомерными БД и хранилищами данных.

– OLAP-сервер – ядро системы, с помощью которого проводится математическая и статистическая обработка многомерных структур данных, и обеспечивается связь между БД и пользователями систем.

– Приложения для работы пользователей, в которых формируются запросы и визуализируются полученные ответы.

Специфика обработки данных OLAP-системами состоит в построении многомерных, то есть имеющих большое количество связей между отдельными элементами, массивов информации. Для формирования таких массивов OLAP-

система собирает данные из различных источников (например, из хранилищ данных, из информационных систем управления предприятием (Enterprise Resource Planning, ERP), из системы взаимодействия с клиентами (Customer Relationship Management, CRM) или через внешний ввод. После этого информация обрабатывается на OLAP сервере и передается в пользовательские приложения.

Хранение и обработка данных с применением OLAP-систем могут осуществляться:

– Непосредственно на рабочих местах пользователей в локальных хранилищах. Такая структура OLAP системы имеет существенные недостатки, связанные со скоростью обработки данных, защищенностью данных и ограниченным применением многомерного анализа.

– В форме реляционных баз данных – при совместной работе OLAP систем с ERP- или CRM-системами. Данные хранятся на сервере этих систем в виде реляционных БД или хранилищ данных. OLAP-сервер обращается к этим базам данных для формирования необходимых многомерных структур и проведения анализа.

– В форме многомерных хранилищ данных на обособленных серверах; данные организованы в виде специального хранилища данных на выделенном *сервере данных*, который преобразует исходные данные в многомерные структуры (OLAP-кубы). Источниками данных для формирования OLAP-куба являются реляционные базы данных ERP- или CRM-систем и/или клиентские файлы. Сервер данных осуществляет предварительную подготовку и обработку данных. OLAP-сервер работает только с OLAP-кубом и не имеет непосредственного доступа к источникам данных (реляционным базам данных, клиентским файлам и др.).

В зависимости от метода хранения и обработки данных все OLAP системы могут быть разделены на три основных вида: реляционные, многомерные и

1. ROLAP (Relational OLAP – реляционные OLAP-системы) – этот вид OLAP-системы работает с реляционными базами данных. Обращение к данным осуществляется напрямую в реляционную базу данных. Данные хранятся в виде реляционных таблиц. Пользователи имеют возможность осуществлять многомерный анализ как в традиционных OLAP-системах. Это достигается за счет применения инструментов SQL и специальных запросов. Одним из преимуществ ROLAP является возможность более эффективно осуществлять обработку большого объема данных. Другим преимуществом ROLAP является возможность эффективной обработки как числовых, так и текстовых данных. К недостаткам ROLAP относится низкая производительность (по сравнению с традиционными OLAP системами), т.к. обработку данных осуществляет сервер OLAP. Другим недостатком является ограничение функциональности из-за применения SQL.

2. MOLAP (Multidimensional OLAP – многомерные OLAP-системы) – этот вид OLAP-систем относится к традиционным системам. Отличие традиционной OLAP-системы, от других систем, заключается в предварительной подготовке и оптимизации данных. Эти системы, как правило, используют выделенный сервер, на котором осуществляется предварительная обработка данных. Данные формируются в многомерные массивы – OLAP-кубы. MOLAP-системы являются самыми эффективными при обработке данных, т.к. они позволяют легко реорганизовать и структурировать данные под различные запросы пользователей. Аналитические инструменты MOLAP позволяют выполнять сложные расчеты. Другим преимуществом MOLAP является возможность быстрого формирования запросов и получения результатов. Это обеспечивается за счет предварительного формирования OLAP кубов. К недостаткам MOLAP системы относится ограничение объемов обрабатываемых данных и избыточность данных, т.к. для формирования многомерных кубов, по различным аспектам, данные приходится дублировать.

3. HOLAP (Hybrid OLAP – гибридные OLAP системы) – эти системы представляют собой объединение систем ROLAP и MOLAP. В гибридных системах объединены преимущества двух систем: поддержка многомерных структур и преимущества реляционных баз данных. HOLAP-системы позволяют хранить большое количество данных в реляционных таблицах, а обрабатываемые данные размещаются в предварительно построенных многомерных OLAP-кубах. Преимущества этого вида систем заключаются в масштабируемости данных, быстрой обработке данных и гибком доступе к источникам данных.

По способам реализации и аппаратным платформам, на которых разворачивается OLAP-система можно выделить:

– Web OLAP – системы с поддержкой web интерфейса;

– Desktop OLAP – системы, БД которых загружена на рабочее место пользователя для организации локальной работы;

– MobileOLAP – системы, работающие с базой данных удаленно, с использованием мобильных устройств;

– Spatial OLAP – системы обработки пространственных данных; появились как результат интеграции географических информационных систем с OLAP-системами. Spatial OLAP-системы обрабатывать данные в буквенно-цифровом формате и в виде визуальных объектов и векторов.

К достоинствам OLAP-систем можно отнести:

– возможность проследить источник информации и определить логическую связь между полученными результатами и исходными данными.

– проведение многовариантного анализа и на основе множества сценариев обработки данных;

– возможность задания требуемой детализации результатов – создаваемые отчеты могут содержать именно ту информацию, которая необходима для принятия решений.

– выявление скрытых зависимостей на основе многомерных связей;

– создание единой платформы для всех процессов анализа и прогнозирования.

Взаимодействуя с OLAP- системой, пользователь может осуществлять гибкий просмотр информации, получать произвольные срезы данных и выполнять аналитические операции детализации, свертки, сквозного распределения, сравнения во времени одновременно по многим параметрам. Вся работа с OLAP-системой происходит в терминах предметной области и позволяет строить статистически обоснованные модели деловой ситуации.

Программные средства OLAP – это инструменты оперативного анализа данных, содержащихся в хранилище. Главной особенностью является то, что эти средства ориентированы на использование не специалистом в области информационных технологий, не экспертом-статистиком, а профессионалом в прикладной области управления – менеджером отдела, департамента, управления, и, наконец, директором. Средства предназначены для общения аналитика с проблемой, а не с компьютером. На рис. 6.3 показан элементарный OLAP-куб, позволяющий производить оценки данных по трем измерениям.

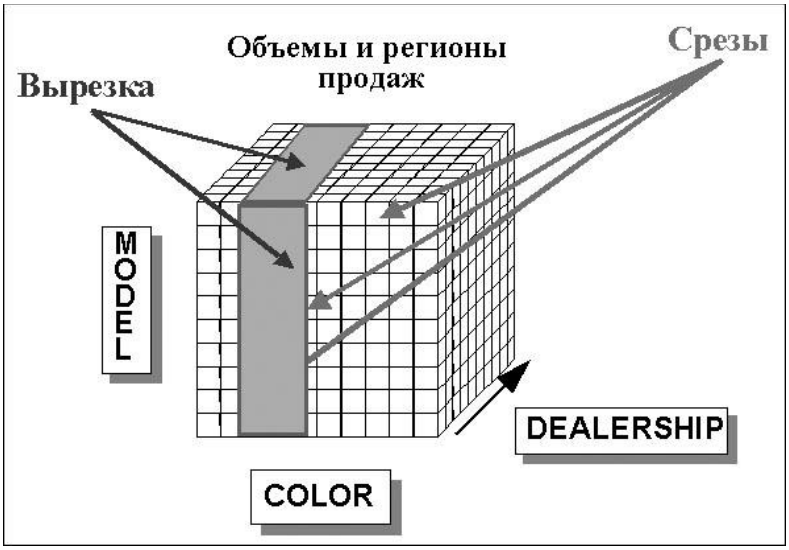

Рисунок 6.3. Элементарный OLAP-куб

Пользователю доступны некоторые стандартные методы анализа, логически следующие из природы OLAP-технологии. В качестве примера рассмотрим методы, применяющиеся для исследования результатов процесса продаж.

Факторный (структурный) анализ. Анализ структуры продаж для выявления важнейших составляющих в интересующем разрезе.

Анализ динамики - выявление трендов, выявление тенденций, сезонных колебаний.

Анализ зависимостей (корреляционный анализ, регрессионный анализ). Сравнение объемов продаж разных товаров во времени для выявления необходимого ассортимента.

Сопоставление (сравнительный анализ) – сравнение результатов продаж во времени, или за заданный период, или для заданной группы товаров.

Дисперсионный анализ – исследование распределения вероятностей и доверительных интервалов рассматриваемых показателей. Применяется для прогнозирования и оценки рисков.

Этими видами анализа возможности OLAP не исчерпываются. Например, применяя в качестве алгоритма вычисления промежуточных и окончательных итогов функции статистического анализа – дисперсию, среднее отклонение, моды более высоких порядков, - можно получить самые изощренные виды аналитических отчетов.

В заключение следует отметить, что повышение быстродействия OLAP-систем при обработке больших объемов данных в настоящее время осуществляется за счет расширения реляционных моделей хранения данных моделями NoSQL. И прежде всего это модели «Ключ – значение» и колоночные модели.

Статистические методы OLAP и Data Mining

Традиционные методы анализа данных (статистические методы) в OLAP в основном ориентированы на проверку заранее сформулированных гипотез и на «грубый» разведочный анализ, составляющий основу оперативной аналитической обработки. Традиционная математическая статистика базируется на концепции 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.).

В то же время одно из основных положений Data Mining – поиск неочевидных закономерностей. Инструменты Data Mining могут находить такие закономерности самостоятельно и также самостоятельно строить гипотезы о взаимосвязях. Поскольку именно формулировка гипотезы относительно зависимостей является самой сложной задачей, преимущество Data Mining по сравнению с другими методами анализа является очевидным. Примеры анализа закономерностей указанных типов при решении задач в различных предметных областях приведены в таблице 8.2.

Таблица 8.2. Примеры формулировок задач при использовании методов OLAP и Data Mining

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

OLAP	Data Mining
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?

Обзор методов Data Mining. Примеры задач.

С помощью методов интеллектуального анализа выполняется систематизация данных по различным критериям количества и качества.

Классификация – прогнозирование одной или нескольких дискретных переменных (меток класса) на основе свойств/атрибутов объектов в наборе данных. Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации определяется набор признаков, которые характеризуют группы объектов – классы; формируется правило отнесения объекта к тому или иному классу на основе имеющегося набора уже размеченных (классифицированных) объектов.

Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; многослойные нейронные сети (neural networks), RBF-сети.

Кластеризация (иерархический и неиерархический) – разделение объектов на группы или кластеры на основе совокупности свойств этих объектов так, чтобы в одну группу попадали объекты со схожими свойствами, а в разных группах находились как можно более отличные друг от друга объекты. Кластеризация является логическим продолжением идеи классификации, её особенность кластеризации заключается в том, что классы объектов изначально не предопределены.

Для решения задачи кластеризации могут применяться методы: деревья решений, k-средних, g-средних, особого вида нейронные сети – сети Кохонена и самоорганизующиеся карты Кохонена.

Байесовские сети – графические структуры, служащие для изображения вероятностных отношений между значительным числом переменных, и структуры, которые необходимы, чтобы осуществить вероятностный вывод, основываясь на представленных переменных.

Поиск явных и неявных взаимосвязей (ассоциаций) – поиск зависимостей между различными событиями или объектами в наборе данных. Наиболее частым применением этого типа алгоритма является создание правил взаимосвязи, которые могут использоваться для анализа потребительской корзины. Поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят

одновременно. Наиболее известный алгоритм решения задачи поиска ассоциативных правил – алгоритм Apriori.

Регрессия – определение одной или нескольких непрерывных числовых переменных на основе других атрибутов в наборе данных.

Анализ последовательностей – отыскание и обобщение часто встречающихся в данных последовательностей действий или ситуаций, таких как серия переходов по веб-сайту или событий, зарегистрированных в журнале перед ремонтом оборудования. Последовательность позволяет найти временные закономерности между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, последовательность определяется высокой вероятностью цепочки связанных во времени событий. Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов (sequential pattern).

Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Таблица 8.3. Примеры задач интеллектуального анализа данных

Примеры задач	Подходящие алгоритмы
Классификация:	
Классификация клиентов из списка потенциальных покупателей как хороших и плохих кандидатов.	Алгоритм дерева принятия решений
Классификация вариантов развития болезней пациентов и исследование связанных факторов	Алгоритм классификации
Вычисление вероятности отказа сервера в течение следующих шести месяцев.	Наивный байесовский алгоритм
Регрессионный анализ:	
Прогноз продаж на следующий год.	Алгоритм временных рядов
Прогноз количества посетителей сайта с учетом прошлых лет и сезонных тенденций.	Алгоритм линейной регрессии
Формирование оценки риска с учетом демографии.	Нейросетевая реализация функции многих переменных
Прогнозирование последовательности:	
Анализ маршрута перемещения по веб-сайту компании.	Алгоритм анализа последовательностей и кластеризации
Анализ факторов, ведущих к отказу сервера.	
Отслеживание и анализ последовательностей действий во время посещения поликлиники с целью формулирования рекомендаций по общим действиям.	
Нахождение групп общих элементов в транзакциях:	
Использование анализа потребительской корзины для определения мест размещения продуктов.	Алгоритм поиска взаимосвязей
Выявление дополнительных продуктов, которые можно предложить купить клиенту.	Алгоритм поиска взаимосвязей
Анализ данных опроса, проведенного среди посетителей события, с целью планирования будущих действий.	Алгоритм дерева принятия решений
Нахождение групп схожих элементов:	Алгоритм кластеризации
Создание профилей рисков для пациентов на основе таких атрибутов, как демография и поведение.	Алгоритм кластеризации

Примеры задач	Подходящие алгоритмы
Анализ пользователей по шаблонам просмотра и покупки. Определение серверов, которые имеют аналогичные характеристики использования.	Алгоритм анализа последовательностей

Согласно *классификации по применяемым стратегиям*, задачи Data Mining подразделяются на следующие группы:

- обучение с учителем;
- обучение без учителя;
- другие.

Категория обучения с учителем представлена следующими задачами Data Mining: классификация, оценка, прогнозирование.

Категория обучения без учителя представлена задачей кластеризации.

В категорию другие входят задачи, не включенные в предыдущие две стратегии.

Инструментарий Data Mining

На рынке программного обеспечения Data Mining существует огромное разнообразие продуктов, относящихся к этой категории. И не растеряться в нем достаточно сложно. Для выбора продукта следует тщательно изучить поставленные задачи, и обозначить те результаты, которые необходимо получить.

Существуют различные варианты решений по внедрению инструментов Data Mining:

- покупка готового программного обеспечения Data Mining;
- покупка программного обеспечения Data Mining, адаптированного под конкретный бизнес;
- разработка Data Mining-продукта на заказ сторонней компанией;
- разработка Data Mining-продукта своими силами;
- различные комбинации вариантов, описанных выше, в том числе использование различных библиотек, компонентов и инструментальные наборы для разработчиков создания встроенных приложений Data Mining.

Сейчас к аналитическим технологиям, в том числе к Data Mining, проявляется огромный интерес. На этом рынке работает множество фирм, ориентированных на создание инструментов Data Mining, а также комплексного внедрения Data Mining, OLAP и хранилищ данных. Инструменты Data Mining во многих случаях рассматриваются как составная часть BI-платформ (Bizness Iintelligence), в состав которых также входят средства построения хранилищ и витрин данных, средства обработки неожиданных запросов (ad-hoc query), средства отчетности (reporting), а также инструменты OLAP.

Разработкой в секторе Data Mining всемирного рынка программного обеспечения заняты как всемирно известные лидеры, так и новые развивающиеся компании. Инструменты Data Mining могут быть представлены либо как самостоятельное приложение, либо как дополнения к основному продукту.

Последний вариант реализуется многими лидерами рынка программного обеспечения. Так, уже стало традицией, что разработчики универсальных статистических пакетов, в дополнение к традиционным методам статистического анализа, включают в пакет определенный набор методов Data Mining. Это такие пакеты как SPSS (SPSS, Clementine), Statistica (StatSoft), SAS Institute (SAS Enterprise Miner). Некоторые разработчики OLAP-решений также предлагают набор методов Data Mining, например, семейство продуктов Cognos. Есть поставщики, включающие Data Mining решения в функциональность СУБД: это Microsoft (Microsoft SQL Server), Oracle, IBM (IBM Intelligent Miner for Data).

Рынок инструментов Data Mining определяется широтой этой технологии и вследствие этого – огромным многообразием программного обеспечения. Приведем классификацию инструментов Data Mining согласно популярному сайту KDnuggets, ведущему информационный бюллетень по искусственному интеллекту, науке о данных и машинному обучению: инструменты общего и специфического назначения; бесплатные и коммерческие инструменты.

Наиболее популярная группа инструментов содержит следующие категории:

универсальные системы:

– наборы инструментов;

специализированные системы:

– инструменты для классификации данных;
– инструменты кластеризации и сегментации;
– инструменты статистического анализа;
– инструменты для анализа текстов (Text Mining), извлечение отклонений (Information Retrieval (IR));
– инструменты визуализации.

Многие специализированные программные продукты совмещают в себе реализацию нескольких методов, в частности, очень часто вместе с основным методом также реализованы и методы визуализации.

Наборы инструментов. К этой категории относятся универсальные инструменты, которые включают методы классификации, кластеризации и предварительной подготовки данных. К этой группе относятся такие известные коммерческие инструменты как:

– Clementine (http://www.spss.com/clementine). Data Mining с использованием Clementine является бизнес-процессом, разработанным для минимизации времени решения задач. Clementine поддерживает процессы:

доступ к данным, преобразования, моделирование, оценивание и внедрение. При помощи Clementine Data Mining выполняется с методологией CRISP-DM.

– DBMiner 2.0 Enterprise (http://www.dbminer.com), мощный инструмент для исследования больших баз данных; использует Microsoft Сервер SQL 7.0 Plato.

– IBM Intelligent Miner for Data (http://www.ibm.com/software/data/iminer/fordata/). Инструмент предлагает последние Data Mining-методы, поддерживает полный Data Mining процесс: от подготовки данных до презентации результатов. Поддержка языков XML и PMML.

– KXEN (Knowledge eXtraction ENgines). Инструмент, работающий на основе теории Вапника (Vapnik) SVM. Решает задачи подготовки данных, сегментации, временных рядов и SVM-классификации.

– Oracle Data Mining (ODM) (http://otn.oracle.com/products/bi/9idmining.html). Инструмент обеспечивает GUI, PL/SQL-интерфейсы, Java-интерфейс. Используемые методы: байесовская классификация, алгоритмы поиска ассоциативных правил, кластерные методы, SVM и другие.

– Polyanalyst (http://www.megaputer.com/). Набор, обеспечивающий всесторонний Data Mining. Сейчас, помимо методов прежних версий, также включает анализ текстов, лес решений, анализ связей. Поддерживает OLE DB for Data Mining и DCOM-технологию.

– SAS Enterprise Miner (http://www.sas.com/). Интегрированный набор, который обеспечивает дружественный GUI. Поддерживается методология SEMMA.

– SPSS (http://www.spss.com/clementine/). Один из наиболее популярных инструментов, поддерживается множество методов Data Mining.

– Statistica Data Miner (http://www.StatSoft.com/). Инструмент обеспечивает всесторонний, интегрированный статистический анализ данных, имеет мощные графические возможности, управление базами данных, а также приложение разработки систем.

Примером российской разработки инструментального набора, кроме Polyanalyst, является пакет ***Loginom***. Loginom – это коммерческая аналитическая low-code платформа, обеспечивающая интеграцию, очистку и анализ данных для принятия эффективных управленческих решений на базе методов визуального проектирования. Является универсальным конструктором с обширным набором готовых компонентов для анализа и исследования данных – от простых математических операций до нейросетей и других методов машинного обучения (ML). Позволяет выстраивать сквозной процесс обработки данных: от ETL-процессов до статистического и интеллектуального анализа данных.

Платформа подходит как для простой аналитики, так и для построения масштабных отказоустойчивых корпоративных систем. Доступна в локальной и серверных версиях, а также в виде облачного сервиса. Пригодна для обработки больших данных. Возможные сферы применения: банковский сектор, ритейл,

логистика, промышленность, маркетинг, сельское хозяйство, телекоммуникации, медицина и т. д. Loginom включен в Реестр российского ПО.

Подходит для среднего бизнеса, корпораций, некоммерческих объединений. Развёртывание на ПК, сервере предприятия, в Облаке. Поддержка русского и английского языков. Включена в Реестр российского ПО. Имеются демо версия и пробная версия.

Oracle Business Intelligence Cloud Service – это онлайн-сервис бизнес-аналитики от компании Oracle, направленный на улучшение качества анализа данных за счёт управления представлениями и визуализаций. Позволяет получать доступ к наиболее актуальным данным, создавать новые аналитические разрезы с нуля или изменять существующие на страницах панели мониторинга для всех сотрудников компании на любом устройстве. Бизнес-пользователи изолированы от сложности представления данных через слой метаданных, который предлагает представление в доступном виде выборок, метрик, группировок OLAP, иерархий и результатов вычислений. Поддерживает создание интерактивных панелей мониторинга и отчеты с широким спектром визуализаций. Позволяет импортировать данные из нескольких источников, независимо от их местоположения, что даёт возможность быстро комбинировать различные данные.

Являясь частью комплексного решения Oracle Analytics, обеспечивает доступ к мобильной аналитике на любом устройстве Android или iOS без дополнительной разработки или настройки, с возможностью работы при отключенной связи (оффлайн). Подходит для среднего бизнеса, корпораций, некоммерческих объединений. Развёртывание на мобильном устройстве или в Облаке. Поддержка русского и английского и еще 9 языков. Демо версия отсутствует. Имеется пробная версия.

SAS Enterprise Miner – это платформа для оптимизации процесса интеллектуального анализа данных при разработке описательных и прогнозных моделей с использованием структурированных алгоритмов и визуальных показателей оценки от компании SAS Enterprise Miner.

Программа позволяет получить информацию, способствующую лучшему принятию решений на основе данных и фактов, оптимизирует процесс интеллектуального анализа данных для быстрой разработки моделей, понимания ключевых взаимосвязей и поиска наиболее важных паттернов в бизнесе.

Подходит для среднего и крупного бизнеса, некоммерческих объединений. Развёртывание на ПК, сервере предприятия или в Облаке. Поддержка английского языка. Имеется демо версия пробная версия.

Polymatica – это коммерческая аналитическая платформа от компании Полиматика Рус для анализа больших объёмов данных из различных предметных областей в интерактивном режиме. Используется как

самостоятельная система и как часть комплексного решения, обеспечивая быструю обработку данных и ad-hoc аналитику.

Высокая скорость взаимодействия обеспечивается за счёт технологий In-Memory и GPU, а также собственной технологии Мультисфер для хранения и сжатия данных. Эта запатентованная технология позволяет пользователям получать доступ к любому узлу данных (мультисферам), содержащему более 1 миллиарда записей, за считанные секунды, что кардинально ускоряет аналитические и исследовательские процессы.

В аналитическую платформу Polymatica встроены методы продвинутой аналитики, такие как кластеризация, прогнозирование, профилирование и ассоциативные правила. Также есть возможность подключить модули машинного обучения через библиотеку Python, адаптированную к API. Подходит для среднего бизнеса, корпораций, некоммерческих объединений. Развёртывание на сервере предприятия, в Облаке (SaaS). Поддерживает русский и английский языки. Включена в Реестр российского ПО. Имеется демо версия. Пробная версия отсутствует.

Orange – свободно-распространяемая аналитическая система с открытым исходным кодом для машинного обучения и визуализации данных, обладающая большим набором исследовательских функций. Разрабатывается Лабораторией биоинформатики Люблянского университета, предназначена для интеллектуального анализа данных (ИАД). Компоненты аналитической платформы называются виджетами, и они варьируются от минималистичной визуализации данных, выбора подмножеств и предварительной обработки до эмпирической оценки алгоритмов обучения и прогностического моделирования.

В рамках визуального программирования аналитические процедуры создаются путём связывания предопределенных или разработанных пользователем блоков (виджетов). Продвинутые пользователи могут использовать Orange в качестве программной библиотеки Python для манипулирования данными и создания новых блоков (виджетов). Подходит для среднего бизнеса, некоммерческих объединений, ИП, специалистов, фрилансеров. Развёртывание на ПК (macOS, Windows, Linux). Поддержка английского языка. Имеются демо версия и пробная версия.

Рынок поставщиков Data Mining активно развивается. Популярность некоторых продуктов возрастает, а некоторых - падает. Это касается как коммерческих, так и свободно распространяемых инструментов. Постоянно появляются новые фирмы-разработчики и новые инструменты, идет постоянная конкурентная борьба за потребителя. Такая конкуренция порождает новые качественные решения. Все большее число поставщиков стремятся объединить в своих инструментах как можно большее число современных методов и технологий, стараясь сокращать отставание существующего программного обеспечения от теоретических разработок в области интеллектуальной обработки данных.