

Аннотация: В данной статье раскрывается понятие технологии интеллектуального анализа данных (Data Mining) как одного из трендов развития систем анализа данных. Определены основные задачи, которые способен выполнить метод Data Mining. Выявлена актуальность данной технологии и ее особенности. Раскрыты проблемы и ограничения в использования интеллектуального анализа данных в современных информационных системах. Проанализирована современная ситуация и тенденции использования технологии интеллектуального анализа данных в различных компаниях. Предложены возможные варианты развития сфер применения метода Data Mining в связи с закономерностями трансформации современного технологического процесса. **Ключевые слова:** интеллектуальный анализ данных, метод Data Mining, выявление неявных и скрытых связей и закономерностей, развитие сферы применения интеллектуального анализа данных, проблемы и ограничения использования интеллектуального анализа данных, обнаружение новых знаний.

Что такое Data Mining

Data Mining (добыча данных, интеллектуальный анализ данных, глубинный анализ данных или просто майнинг данных) — это процесс, используемый компаниями для превращения необработанных больших данных в полезную информацию. Также для этой технологии используется менее популярный термин «обнаружение знаний в данных» или KDD (knowledge discovery in databases).

Если термином Big Data обозначают все большие данные — как обработанные, так и нет, то Data Mining представляет собой процесс глубокого погружения в эти данные для извлечения ключевых знаний.

Автор термина Data Mining Григорий Пятецкий-Шапиро определял его как процесс обнаружения в сырых данных ранее неизвестных, нетривиальных,

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

Используя программное обеспечение для поиска закономерностей в больших пакетах данных, предприятия могут выстраивать маркетинговые стратегии, управлять кредитными рисками, обнаруживать мошенничество, фильтровать спам или даже выявлять настроения пользователей.

Интеллектуальный анализ данных зависит от эффективного сбора, хранения и компьютерной обработки данных. Data Mining считается отдельной дисциплиной в области науки о данных.

Термин «интеллектуальный анализ данных» фигурировал в академических журналах еще в 1970 году, но по-настоящему популярным он стал только в 1990-х после появления интернета. Тогда компаниям потребовалось анализировать большие объемы разнородных данных, чтобы отыскать нетривиальные паттерны и научиться предсказывать поведение клиентов. Обычные модели статистики оказались неспособны справиться с этой задачей.

Первые системы Data Mining предназначались для обработки данных о продажах в супермаркетах по нескольким параметрам, включая их объем по регионам и тип продукта.

Data Mining — это методология и процесс обнаружения в больших массивах данных, накапливающихся в информационных системах компаний, ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining является одним из этапов более масштабной методологии Knowledge Discovery in Databases.

Знания, обнаруженные в процессе Data Mining, должны быть нетривиальными и ранее неизвестными. Нетривиальность предполагает, что такие знания не могут быть обнаружены путем простого визуального анализа. Они должны описывать связи между свойствами бизнес-объектов,

предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и к новым объектам.

Практическая полезность знаний обусловлена возможностью их использования в процессе поддержки принятия управленческих решений и совершенствовании деятельности компании.

Знания должны быть представлены в виде, понятном для пользователей, которые не имеют специальной математической подготовки. Например, проще всего воспринимаются человеком логические конструкции «если, то». Более того, такие правила могут быть использованы в различных СУБД в качестве SQL-запросов. В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду.

Data Mining — это не один, а совокупность большого числа различных методов обнаружения знаний. Все задачи, решаемые методами Data Mining, можно условно разбить на шесть видов:

- Классификация;
- Регрессия;
- Кластеризация;
- Ассоциация;
- Последовательные шаблоны;
- Анализ отклонений.

Data Mining носит мультидисциплинарный характер, поскольку включает в себя элементы численных методов, математической статистики и теории вероятностей, теории информации и математической логики, искусственного интеллекта и машинного обучения.



Рисунок 1 – Data Mining

Почему интеллектуальный анализ данных – это важно?

Интеллектуальный анализ данных является важной частью любой успешной аналитической инициативы. Компании могут использовать процесс обнаружения знаний, чтобы повышать доверие клиентов, находить новые источники дохода и привлекать новых клиентов. Эффективный интеллектуальный анализ данных помогает в различных аспектах бизнес-планирования и управления операциями. Ниже приведены несколько примеров того, как различные отрасли используют интеллектуальный анализ данных.

Телекоммуникации, СМИ и технологии

Такие высококонкурентные отрасли, как телекоммуникации, СМИ и технологии, используют интеллектуальный анализ данных для улучшения обслуживания клиентов путем выявления закономерностей в их поведении. Например, компании могут анализировать модели использования пропускной способности и предоставлять индивидуальные улучшенные услуги или рекомендации.

Банковское дело и страхование

Финансовые службы могут использовать приложения для интеллектуального анализа данных, чтобы решать сложные проблемы, связанные с мошенничеством, соблюдением нормативных требований, управлением рисками и потерей клиентов. Например, страховые компании

могут определять оптимальную стоимость товаров, сравнивая прошлые показатели их эффективности с ценами конкурентов.

Сфера образования

Образовательные учреждения могут использовать алгоритмы интеллектуального анализа данных для проверки знаний учащихся, персонализации уроков и игрофицирования обучения. Единое представление успеваемости учащихся на основе данных может помочь преподавателям лучше понимать потребности учащихся и оказывать им более качественную поддержку.

Обрабатывающая промышленность

Производственные службы могут использовать методы интеллектуального анализа данных для предоставления прогнозной аналитики в режиме реального времени, чтобы повышать общую эффективность оборудования, уровни обслуживания, качество продукции и эффективность цепочки поставок. Например, производители могут использовать архивные данные для прогнозирования износа производственного оборудования и планирования технического обслуживания. В результате они могут оптимизировать производственные графики и сократить время простоя.

Розничная торговля

Розничные компании имеют большие клиентские базы данных с необработанными данными о поведении покупателей. С помощью интеллектуального анализа данных можно обрабатывать эти данные для получения соответствующей информации для маркетинговых кампаний и прогнозов продаж. Благодаря более точным моделям данных розничные компании могут оптимизировать продажи и логистику, что позволит повысить удовлетворенность клиентов. Например, с помощью интеллектуального анализа данных можно выявить популярные сезонные товары, которыми можно запастись заранее, чтобы избежать дефицита.

Задачи Data Mining

Модели интеллектуального анализа данных применяются для нескольких типов задач:

прогнозирование: оценка продаж, предсказание нагрузки сервера или его времени простоя;

риск и вероятность: выбор подходящих заказчиков для целевой рассылки, определение точки баланса для рискованных сценариев, назначение вероятностей по диагнозам или другим результатам;

рекомендации: определение продуктов, которые будут продаваться вместе, создание рекомендательных сообщений;

поиск последовательностей: анализ выбора заказчиков во время совершения покупок, прогнозирование их поведения;

группирование: разделение заказчиков или событий на кластеры, анализ и прогнозирование общих черт этих кластеров.



Рисунок 2 – Задачи data mining

Программы для Data Mining

Существует множество программ, которые могут выполнять задачи Data Mining. Вот некоторые примеры.

SAS Enterprise Miner — набор методов интеллектуального анализа данных, который применяется для решения таких задач, как обнаружение случаев мошенничества, минимизация финансовых рисков, оценка и

прогнозирование потребностей в ресурсах, повышение эффективности маркетинговых кампаний и снижение оттока клиентов. Имеет удобный и понятный интерфейс, позволяющий пользователям самостоятельно создавать модели анализа и прогнозирования. Показывает высокую производительность даже при работе с огромным массивом разрозненных данных.

Microsoft Analysis Services — предназначен для приложений бизнес-аналитики, анализа данных и создания отчетов. Службы доступны на разных платформах, в том числе на облаке Azure. Предусмотрен механизм для создания собственных алгоритмов и добавления их в качестве новой функции интеллектуального анализа данных.

SAS Customer Intelligence 360 — это платформа, которая позволяет бизнесу планировать и реализовывать маркетинговые кампании, анализировать их итоги и отслеживать потоки клиентов. Она в реальном времени собирает подробную информацию о действиях клиентов на веб-страницах, в том числе анонимных пользователей, учитывая контекст. Затем платформа дает рекомендации о времени и месте размещения контента на страницах и в мобильных приложениях для конкретного клиента.



Рисунок 3 – Методика построения модели интеллектуального анализа данных

Какие существуют типы интеллектуального анализа данных?

В зависимости от данных и цели анализа интеллектуальный анализ данных может иметь различные направления или специализации. Рассмотрим некоторые из них ниже.

Анализ процессов

Анализ процессов – направление интеллектуального анализа данных, целью которого является обнаружение, мониторинг и улучшение бизнес-процессов. Анализ процессов подразумевает извлечение знаний из журналов событий, доступных в информационных системах. Это позволяет организациям видеть и понимать, что происходит в этих процессах изо дня в день.

Например, в компаниях, работающих в сфере электронной торговли, выполняется множество процессов, в частности закупка, продажа, оплата, сбор и доставка товаров. Проанализировав журналы данных о закупках, компании могут увидеть, что надежность поставок их поставщиков составляет 54 % или что 12 % поставщиков постоянно доставляют товары раньше срока. Эту информацию они могут использовать для оптимизации отношений с поставщиками.

Анализ текста

Для интеллектуального анализа текста или текстовых данных используется специальное программное обеспечение для чтения и понимания текста. Специалисты по работе с данными используют интеллектуальный анализ текста для автоматизации поиска знаний в письменных ресурсах, таких как веб-сайты, книги, электронные письма, отзывы и статьи.

Например, цифровая медиа-компания может использовать интеллектуальный анализ текста, чтобы автоматически читать комментарии к своим онлайн-видео и классифицировать отзывы зрителей как положительные или отрицательные.

Прогностический анализ

Прогностический интеллектуальный анализ данных использует бизнес-аналитику для прогнозирования тенденций. Он позволяет руководителям бизнеса изучать влияние их решений на будущее своих компаний и принимать более эффективные решения.

Например, компания может изучить данные о прошлых возвратах продукции, чтобы разработать схему гарантии, которая поможет избежать убытков. С помощью прогностического анализа компании могут предсказать потенциальное количество возвратов в следующем году и создать годовой гарантийный план, который учитывает убытки при определении цены товара.



Рисунок 4 – Интеллектуальный анализ данных

Вывод:

Рынок систем Data Mining растет. Этому способствует деятельность крупных корпораций: SAS, IBM, Microsoft, Oracle и других. Ожидается, что к 2027 году объем глобального рынка расширенной аналитики вырастет на 23,1% и достигнет отметки в \$56,2 млрд.

Последние тенденции в Data Mining включают развитие методов анализа с элементами виртуальной и дополненной реальности, их интеграцию с системами баз данных, добычу биологических данных для инноваций в медицине, веб-майнинг (анализ данных в интернете), анализ данных в реальном времени, а также меры по защите конфиденциальности при добыче

данных. Лидеры отрасли считают, что в будущем майнинг данных будет применяться в интеллектуальных приложениях, которые будут встроены в корпоративные хранилища данных.

Главной проблемой обнаружения закономерностей в данных является время, которое требуется для перебора информационных массивов. Известные методы либо искусственно ограничивают такой перебор, либо строят целые деревья решений, которые снижают эффективность поиска. Решение этой проблемы остается главной целью разработчиков продуктов для Data Mining.

Список литературы:

Баклин Н. Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям (+CD). — СПб.: Изд. Питер, 2009. — 624 с.

Дюк В., Самойленко А. Data Mining: учебный курс (+CD). — СПб.: Изд. Питер, 2001. — 368 с.

Журавлёв Ю.И., Рязанов В.В., Сенько О.В. РАСПОЗНАВАНИЕ. Математические методы. Программная система. Практические применения. — М.: Изд. «Фазис», 2006. — 176 с. — ISBN 5-7036-0108-8.

Зиновьев А. Ю. Визуализация многомерных данных. — Красноярск: Изд. Красноярского государственного технического университета, 2000. — 180 с.

Чубукова И. А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. — 382 с. — ISBN 5-9556-0064-7.

Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг): Навч. посібник. — К.: КНЕУ, 2007. — 376 с.

Ian H. Witten, Eibe Frank and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664. — ISBN 9780123748560.

Дэйв Миллнер, Надим Хан. HR-аналитика. Практическое руководство по работе с персоналом на основе больших данных = Introduction to People

Analytics: A practical guide to data-driven HR. — М.: Альпина Паблишер , 2022. — 384 с. — ISBN 978-5-9614-7831-0.

Орлов А.И. Искусственный интеллект: статистические методы анализа данных : учебник. — М.: Ай Пи Ар Медиа, 2022. — 843 с. — ISBN 978-5-4497-1470-1 [1]

Орлов А.И., Луценко Е.В. Анализ данных, информации и знаний в системной нечеткой интервальной математике: научная монография. — Краснодар: КубГАУ, 2022. — 405 с. [2]

А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. — Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург, 2004. — 336 с.

А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. — Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-е изд., — СПб.: БХВ-Петербург, 2007. — 384 с.

А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. — Анализ данных и процессов: учеб. пособие. 3-е изд., — СПб.: БХВ-Петербург, 2009. — 512 с.