УДК-004.8

Арсентьева Н.В., магистрант кафедры прикладной информатики и информационных технологий
Скрипин А.А, аспирант кафедры прикладной информатики и информационных технологий
Скрипина И. И., старший преподаватель кафедры прикладной информатики и информационных технологий
Белгородский государственный национальный исследовательский университет, г. Белгород

СОВРЕМЕННЫЕ И ПЕРСПЕКТИВНЫЕ АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Аннотация: Данная статья посвящена обзору современных и перспективных алгоритмов интеллектуального анализа данных (ИАД). В работе рассматриваются ключевые алгоритмы классификации, кластеризации, поиска ассоциативных правил, анализа временных рядов, снижения размерности и обработки текста. Авторы описывают области применения ИАД в различных сферах, включая бизнес, здравоохранение, финансы и производство. Особое внимание уделяется перспективам развития алгоритмов ИАД, таким как глубокое обучение, обработка естественного языка, обучение с подкреплением и автоматизация процесса обучения моделей. Статья предоставляет комплексный обзор текущего состояния и будущих направлений развития в области интеллектуального анализа данных.

Ключевые слова: *интеллектуальный анализ данных, ассоциативные правила, алгоритмы классификации*

Arsentieva N.V., Undergraduate student of the Department of Applied Informatics and Information Technology

Skripin A.A., post-graduate student of the Department of Applied Informatics and Information Technology
I. I. Skripina, Senior Lecturer at the Department of Applied Informatics and Information Technology
Belgorod State National Research University, Belgorod

MODERN AND PROMISING DATA MINING ALGORITHMS

Abstract: This article is devoted to an overview of modern and promising algorithms for data mining (IAD). The paper discusses key algorithms for classification, clustering, search for associative rules, time series analysis, dimensionality reduction and text processing. The authors describe the application areas of IAD in various fields, including business, healthcare, finance and manufacturing. Special attention is paid to the prospects for the development of IAD algorithms, such as deep learning, natural language processing, reinforcement learning and process automation

В эпоху трансформаций и экспоненциального роста объемов данных интеллектуальный анализ данных (ИАД) становится ключевым фактором для извлечения ценной информации и поддержки принятия решений в различных видах деятельности. Современные алгоритмы ИАД позволяют обрабатывать большие массивы структурированных и неструктурированных данных, выявлять скрытые закономерности и обеспечивать действенные идеи для бизнеса, науки и общества.

Под интеллектуальным анализом данных понимают обработку информации и выявление в ней тенденции, которая помогает принимать решения. Существует множество различных методов интеллектуального анализа данных, моделирования запросов обработки и сбора информации [Data mining: страница Википедии об интеллектуальном анализе данных. [3]

В настоящее время объем предоставляемых данных безграничен, однако часто этот обширный массив информации замаскирован под

плотными покровами комплексности. Организации, деятельность которых охватывает разнообразные области, постоянно стремятся раскрывать ценные аспекты из масштабных объемов предоставляемой информации. Этот стремительный поиск информации ускорил процесс развития интеллектуального анализа данных — области, предоставляющей возможность предприятиям, научным исследователям и принимающим решения лицам выявлять скрытые закономерности и тенденции, стимулирующие инновационные процессы и служащие основой для принятия ключевых решений. [4]

Рис. 1. Процесс цифровой трансформации в России [4]

Обзор классических и современных алгоритмов интеллектуального анализа данных, широко применяемых в различных областях.

1. Алгоритмы классификации:

— Случайный лес (Random Forest): Ансамбльный метод, основанный на построении последовательного решения. Отличается высоким авторитетом и стойкостью к переобучению;

— Градиентный бустинг (Gradient Boosting): Семейство алгоритмов, включающее XGBoost, LightGBM и CatBoost. Эффективны для решения задач с большим количеством признаков и функцией зависимости;

— Глубокие нейронные сети (Deep Neural Networks): Многослойные структуры, способные к автоматическому извлечению признаков. Особые эффективны в задачах компьютерного зрения и обработки естественного языка;

Задача классификации — определить к какому классу относятся те или иные данные; при этом множество классов к одному из которых впоследствии можно отнести исследуемый объект заранее известно. (2)

2. Алгоритмы кластеризации:

— DBSCAN (Пространственная кластеризация приложений с шумом на основе плотности): Алгоритм, способ нахождения кластеров последовательной формы и выявлять выбросы;

— HDBSCAN (Hierarchical DBSCAN): Улучшенная версия DBSCAN, автоматически определяющая оптимальные параметры кластеризации;

— Спектральная кластеризация: Метод, основанный на анализе натуральных векторов матриц, подобных данным. Эффективен для сложных структур данных.

Задача классификации — определить к какому классу относятся те или иные данные; при этом множество классов к одному из которых впоследствии можно отнести исследуемый объект заранее известно. [2]

3. Алгоритмы поиска ассоциативных правил:

— FP-Growth (частый шаблон роста): эффективный алгоритм для поиска часто встречающихся наборов элементов без генерации критериев;

— ECLAT (преобразование класса эквивалентности): алгоритм, использующий вертикальное представление данных для быстрого поиска ассоциаций.

Ассоциативные правила позволяют находить закономерности между связанными событиями.

4. Алгоритмы анализа временных рядов:

— ARIMA (авторегрессионное интегрированное скользящее среднее): Классический метод для определения и прогнозирования временных рядов;

— Пророк: Разработанный алгоритм Facebook для прогнозирования временных рядов с учетом сезонности и праздников;

— LSTM (Long Short-Term Memory): архитектура рекуррентных нейронных сетей, эффективная для постоянного анализа данных.

Один из наиболее важных инструментов в аналитике данных является анализ временных рядов. Временной ряд - это последовательность наблюдений за определенным параметром в разные моменты времени. Таким образом, временной ряд содержит информацию о том, как изменяется параметр со временем и является одним из важных компонентов современной аналитики данных и имеет большие практические применения в различных областях.

5. Алгоритмы снижения размерности:

— t-SNE (t-распределенное стохастическое встраивание соседей): Нелинейный метод визуализации многомерных данных, сохраняющий локальную структуру;

— UMAP (аппроксимация и проекция равномерного многообразия): алгоритм, обеспечивающий быстрое снижение размерности с сохранением глобальной структуры данных.

Уменьшение размерности широко используется в области машинного обучения и анализа данных. Его цель состоит в том, чтобы упростить

обработку данных за счет уменьшения количества объектов в наборе при сохранении ключевой информации.

6. Алгоритмы обработки текста:

— BERT (представления двунаправленного кодировщика от трансформаторов): модель глубокого обучения для задач обработки естественного языка, использующая механизм внимания;

— Word2Vec: метод создания векторных представленных слов, позволяющий захватывать семантические отношения.

Обработка текста помогает предприятиям автоматизировать процессы и получать ценную информацию из данных. Современные алгоритмы ИАД характеризуются высокой производительностью, работают с определяемыми объемами данных и адаптируются к задачам любого типа. Многие из них интегрируются в комплексные системы анализа данных, создавая синергетический эффект при решении простых аналитических задач.

Области применения технологий и методов, использующих алгоритмы интеллектуального анализа данных, действительно находят широкое применение в различных сферах:

— Бизнес и маркетинг: сегментация клиентов, прогнозирование спроса, анализ потребительской корзины, оптимизация цепочек поставок, выявление мошенничества;

— Здравоохранение: диагностика заболеваний, персонализированная медицина, анализ медицинских изображений, прогнозирование эпидемий, оптимизация работы медицинских учреждений;

— Финансы: оценка кредитных рисков, прогнозирование движения цен на рынке, выявление аномальных транзакций, портфельная оптимизация, автоматизированная торговля;

— Производство: предиктивное обслуживание оборудования, контроль качества продукции, оптимизация производственных процессов, управление запасами;

— Телекоммуникации: анализ поведения пользователей, прогнозирование оттока клиентов, оптимизация сетевой инфраструктуры, выявление аномалий в работе сети;

— Транспорт и логистика: оптимизация маршрутов, прогнозирование трафика, управление автопарком, анализ поведения водителей;

— Образование: персонализация обучения, прогнозирование успеваемости студентов, оптимизация учебных программ, анализ эффективности образовательных методик;

— Энергетика: прогнозирование потребления энергии, оптимизация работы энергосетей, предсказание сбоев в оборудовании, анализ данных с умных счетчиков;

— Государственное управление: выявление налоговых махинаций, анализ эффективности государственных программ, прогнозирование социально-экономических показателей, оптимизация городской инфраструктуры;

— Научные исследования: анализ больших массивов экспериментальных данных, моделирование сложных систем, поиск закономерностей в геномных данных, обработка данных с научных приборов и сенсоров.

К примеру, Сбербанк предоставляет сервис «Сбор Аналитики», основанный на анализе денежных потоков, продаж товаров и других параметров, предоставляя данные по отраслям рынка или регионам. Как компании, так и государственные органы могут использовать этот инструмент для оценки потенциала развития региона, а российская сеть «Лента» провела анализ данных более чем 90% держателей карт лояльности, выделив сегменты по покупательскому поведению. Это позволило оптимизировать ассортимент, управлять выкладкой и ценами. Амазон в свою очередь предоставил продавцам доступ к данным о текущих запросах

покупателей, упрощая процесс выбора продуктов для продажи. [© 2024 ООО «ИЗДАТЕЛЬСТВО СК ПРЕСС» https://www.itweek.ru/bigdata/article/detail.php?ID=229390]

Перспективы развития алгоритмов интеллектуального анализа данных.

На основе общих тенденций в области науки о данных и искусственного интеллекта можно выделить следующие перспективы развития алгоритмов интеллектуального анализа данных:

1. Глубокое обучение (Deep Learning): Развитие более сложных нейронных сетей и алгоритмов глубокого обучения для решения различных задач, таких как обработка изображений, распознавание речи, семантический анализ и многое другое.

2. Обработка естественного языка (Natural Language Processing, NLP): Усовершенствование алгоритмов NLP для более точного анализа и понимания текстовой информации, включая машинный перевод, сентимент-анализ, суммаризацию текста и другие приложения.

3. Обучение с подкреплением (Reinforcement Learning): Развитие методов обучения с подкреплением для обучения агентов принимать оптимальные решения в условиях неопределенности, что может применяться в автономных системах, играх и других областях.

4. Объединение различных типов данных (Structured, Unstructured, Semi-structured): Возможность создания алгоритмов, способных работать с различными типами данных, включая структурированные таблицы, изображения, видео, аудио, текст и т. д.

5. Автоматизация процесса обучения модели (AutoML): Развитие технологий AutoML для автоматизации процесса выбора, настройки и оптимизации моделей машинного обучения без необходимости глубоких знаний эксперта.

6. Интерпретируемость моделей (Interpretable Machine Learning): Улучшение методов объяснения и визуализации результатов работы моделей машинного обучения для повышения доверия пользователей и обеспечения прозрачности алгоритмов.

7. Развитие гибридных моделей (Hybrid Models): Комбинирование различных алгоритмов и моделей для решения сложных задач и повышения качества прогнозов и классификации.

8. Федеративное обучение (Federated Learning): Развитие методов федеративного обучения, который позволяет обучать модели на распределенных данных с помощью совместного обучения и обмена минимальными количествами информации.

Эти направления исследования и развития алгоритмов интеллектуального анализа данных являются ключевыми для эффективного использования данных в различных областях, с целью получения ценной информации, выявления закономерностей и принятия обоснованных решений.

Вывод.

Интеллектуальный анализ данных является критически важным инструментом в эпоху цифровой трансформации и экспоненциального роста объемов информации. Рассмотренные в статье современные алгоритмы ИАД демонстрируют высокую эффективность в решении широкого спектра задач, от классификации и кластеризации до анализа временных рядов и обработки текста. Их применение в различных отраслях, от бизнеса и финансов до здравоохранения и государственного управления, свидетельствует о универсальности и важности этих методов.

Перспективы развития алгоритмов ИАД связаны с дальнейшим совершенствованием методов глубокого обучения, обработки естественного языка, обучения с подкреплением и автоматизации процессов машинного обучения. Особое внимание уделяется развитию интерпретируемых моделей

и федеративного обучения, что отражает растущую потребность в прозрачности алгоритмов и защите конфиденциальности данных.

В целом, интеллектуальный анализ данных продолжает развиваться, адаптируясь к новым вызовам и потребностям цифровой экономики. Будущее этой области лежит в интеграции различных подходов и создании более гибких, эффективных и этичных алгоритмов, способных извлекать ценные знания из растущих объемов сложных и разнородных данных, что усилит аналитические возможности и повысит эффективность принимаемых на их основе решений.

Литература

1. Дюк, В. . Data Mining. Учебный курс / В. Дюк, А. Самойленко. - Санкт-Петербург : Питер, 2001. - 366 с.

2. Певченко, С. С. Методы интеллектуального анализа данных / С. С. Певченко. — Текст : непосредственный // Молодой ученый. — 2015. — № 13 (93). — С. 167-169. — URL: https://moluch.ru/archive/93/20875/

3. Степанов, Р. Г. Технология Data Mining: Интеллектуальный Анализ Данных / Р. Г. Степанов. — 2008 : - Казанский Государственный Университет им. И.Ульянова-Ленина, 2008. — 546 с.

4. Мокшанов, Михаил / Михаил Мокшанов. — Текст : электронный // : [сайт]. — URL: https://www.itweek.ru/bigdata/article/detail.php?ID=229390&ysclid=lxujisdis644865546 (дата обращения: 25.06.2024).