УДК 004.65+004.85

BIG DATA И МАШИННОЕ ОБУЧЕНИЕ

Н.И. Потапенко
Старший преподаватель кафедры инженерной психологии и эргономики

В.А. Буд-Гусаим
Студент кафедры инженерной психологии и эргономики

А.Н. Василькова
Ассистент кафедры инженерной психологии и эргономики, магистр
a.vasilkova@bsuir.by

Н.И. Потапенко
Старший преподаватель кафедры инженерной психологии и эргономики Белорусского государственного университета информатики и радиоэлектроники.

А.А. Войтович
Студент кафедры инженерной психологии и эргономики Белорусского государственного университета информатики и радиоэлектроники.

А.Н. Василькова
Ассистент кафедры инженерной психологии и эргономики Белорусского государственного университета информатики и радиоэлектроники, магистр.

Аннотация. Большие данные (Big Data) является мощным инструментом для решения и переосмысления проблем в IT сфере, бизнесе, медицине и т.д. Поскольку объем данных продолжает расти с невероятной скоростью, способность правильно их обрабатывать стала важнейшим ключом к созданию богатого источника для обучения алгоритмов.Для работы с Big Data используют различные методы машинного обучения (ML), такие как деревья решений, алгоритмы кластеризации, нейросети, которые способны учиться на обрабатываемых ими данных. Цель статьи – дать представление о том, как организации могут использовать ML и Big Data, чтобы получить конкурентное преимущество в современном мире, управляемом данными.

Ключевые слова: большие данные, машинное обучение, алгоритм, нейросеть, обработка наборов данных, зависимость, самообучение.

Введение.

В современную цифровую эпоху объем генерируемых данных растет экспоненциально, что создает требования наличия мощных и эффективных методов обработки данных. Однако из-за объема и сложности Big Data традиционные методы обработки оказались неточными и неэффективными. Здесь на помощь приходит ML, предоставляющее передовые алгоритмы, которые автоматически выявляют закономерности, делают прогнозы и учатся на основе данных.

Использование ML для обработки Big Data практикуется все чаще с каждым днем. Объединив эти две технологии, организации могут анализировать свои данные, принимать более обоснованные решения и повышать общую эффективность своей работы. Кроме того, самообучающиеся алгоритмы ML постоянно развиваются. Объединяя Big Data и ML, алгоритмы Machine Learning помогают нам справляться с непрерывным потоком данных, а объемы и разнообразие потоков данных прокачивают алгоритмы, делая их более совершенными. В статье рассматриваются проблемы и потенциал этого подхода, приводятся примеры его применения в различных областях и дается обзор последних исследований в этой области. Цель статьи – дать представление о том, как организации могут использовать ML и Big Data, чтобы получить конкурентное преимущество в современном мире, управляемом данными.

Основная часть.

Big Data- это коллекция различных сложных наборов данных. С каждым днем их становится все больше и больше, в результате, их трудно хранить и обрабатывать. Чтобы Bid Data превратилась в полезный инструмент для человека, ее обработка должна включать в себя сбор, отбор, хранение, поиск, совместное использование, передачу, анализ и визуализацию этих данных. Работа с большими данными строится вокруг пяти основных принципов (с англ. V's of Big Data: Volume, Velocity, Variety, Veracity, Value):

Объем: количество собираемой компаниями информации действительно огромно, поэтому ее объем становится критическим фактором в аналитике. Скорость: практически все происходящее вокруг нас (поисковые запросы, социальные сети и т. д.) очень быстро генерирует новые данные, многие из которых могут быть использованы в бизнес-решениях.

Разнообразие: генерируемая информация неоднородна и может быть представлена в различных форматах, таких, например, как видео, текст, базы данных, числа, сенсорные данные и т. д. Понимание типов больших данных является ключевым фактором для раскрытия их ценности. Достоверность: данные высокой достоверности содержат много записей, которые ценны для анализа и которые вносят значимый вклад в общие результаты. Данные с низкой достоверностью содержат высокий процент бессмысленной информации, которая называется шумом.

Ценность: возможность трансформировать большие данные в бизнес-решения.

Big Data пополняется из большого количества источников. Основные из них:

– Социальные сети
– Облачные хранилища
– Веб-сайты
– Интернет Вещей (IoT)

Machine Learning (ML) – это форма искусственного интеллекта, которая позволяет анализировать данные, принимать на их основе решения, перекрывать некоторые проблемы в сфере работы, автоматизируя большую часть работы с информацией. Он работает путем изучения данных и выявления закономерностей и требует минимального вмешательства человека. Главной особенностью ML является его возможность самообучаться с помощью обработанных данных. Почти любую задачу, которую можно выполнить с помощью шаблона или набора правил, определяемых данными, можно автоматизировать с помощью машинного обучения. Это позволяет компаниям трансформировать процессы, которые ранее были доступны только людям, например, ответы на звонки в службу поддержки клиентов, ведение бухгалтерского учета и просмотр резюме.

Использование алгоритмов ML для анализа Big Data - логичный шаг для компаний, стремящихся максимизировать потенциальную ценность своих данных. Они могут быть реализованы во всех элементах работы с Big Data, включая маркировку и сегментацию данных, анализ данных и моделирование сценариев. Алгоритмы учатся на данных по мере их обработки, в отличие от традиционных аналитических систем, основанных на правилах, которые следуют четким инструкциям.

Примеры использования Big Data и ML. Сегментация. Для любого бизнеса одним из важнейших ключей для успеха является целевая аудитория. Каждое предприятие должно понимать аудиторию и рынок, на которые оно хочет ориентироваться. Вот почему предприятиям необходимо проводить исследования рынка, которые могут глубоко проникнуть в мысли потенциальных клиентов и предоставить ценные данные. С этим может помочь ML, использующее контролируемые и неконтролируемые алгоритмы для точной интерпретации потребительских моделей и поведения. Средства массовой информации и индустрия развлечений используют ML, чтобы понимать, что нравится и не нравится их аудитории, и нацеливать на нее нужный контент. Например, Netflix, Facebook, Google, Twitter и т.д.

Изучение поведения клиентов. ML также помогает компаниям исследовать поведение

аудитории и создавать прочную структуру своих клиентов. Эта система ML, известная как пользовательское моделирование, является прямым результатом взаимодействия человека с компьютером. Он собирает данные, чтобы захватить мысли пользователя и позволить бизнес-предприятиям принимать разумные решения. Amazon, Uber, LinkedIn и другие полагаются на системы моделирования пользователей, чтобы знать своих пользователей наизнанку и делать соответствующие предложения.

Предсказание трендов. Алгоритмы ML используют Big Data для изучения будущих тенденций и прогнозирования их для бизнеса. С помощью этих данных ML может самостоятельно учиться новому и улучшать свои аналитические навыки каждый день. Таким образом, он не просто вычисляет данные, но ведет себя как интеллектуальная система, которая использует прошлый опыт для формирования будущего. Бренд кондиционеров может полагаться на ML, чтобы прогнозировать спрос на кондиционеры в следующем сезоне и соответствующим образом планировать свое производство.

Помощь в принятии решений. ML использует метод, называемый анализом временных рядов, который позволяет анализировать массив данных вместе. Это отличный инструмент для сбора и анализа данных, облегчающий менеджерам принятие решений на будущее. Предприятия, особенно розничные торговцы, могут использовать этот метод, усиленный машинным обучением, для предсказания будущего с похвальной точностью.

Как это работает?

Структура использования ML с Big Data (MLBiD) состоит из 5 компонентов: Machine Learning, Big Data, пользователь, система, сфера использования. Все эти компоненты взаимодействуют между собой обоюдно. Так, например, Big Data служит «материалом» для ML, оно же генерирует новые данные, которые становятся частью Big Data.При работе с данными ML проходит 3 этапа: предварительная обработка, обучение, оценка. Предварительная обработка. Она преобразует исходные данные в соответствующую для последующих этапов обучения форму. Исходные данные, скорее всего, неструктурированные, неполные непоследовательные и содержат много «шума». Этап предварительной обработки преобразует такие данные посредством очистки, извлечения, преобразования и объединения данных.

Обучение. На этапе обучения выбирается и настраивается алгоритм машинного обучения для получения желаемых результатов с использованием предварительно обработанных входных данных. Некоторые модели обучения, в частности репрезентативное обучение, также могут быть использованы для предварительной обработки данных. Во время обучения алгоритм изучает закономерности и отношения в данных.Оценка. Алгоритм и его работу необходимо оценить, чтобы определить точность и производительность. Это предполагает тестирование модели на отдельном наборе данных, который не использовался во время обучения. Результаты оценки могут привести к корректировке параметров выбранных алгоритмов обучения и/или выбору других алгоритмов. Однако работая с Big Data объем данных настолько велик, что обрабатывать в памяти все сразу не является возможным. В таких случаях используют алгоритмы, методы которых умеют работать с слишком большими данными. Один из подходов заключается в использовании сред распределенных вычислений, таких как Apache Spark или Hadoop, которые позволяют обрабатывать данные параллельно на нескольких машинах.

Другой подход заключается в использовании метода, называемого «онлайн-обучение», при котором модель постоянно обновляется по мере поступления новых данных. Это может быть полезно при работе с динамическими данными.

Кроме того, существуют такие методы, как извлечение признаков и уменьшение размерности, которые можно использовать для уменьшения размера данных при сохранении важной информации.

Плюсы и минусы.

Минусы:

Дубликаты – на просторах Big Data есть много дублированной информации. Это

происходит, когда несколько выборок имеют одинаковые сущности.

Маркировка – из-за большого количества данных и их источников сложно маркировать наборы данных правильным образом, чтобы при поиске определенной информации не возникало проблем и временных затрат. В результате нужный набор данных может не охватывать все специфические для пользователя контексты.

Хранение - работа с Big Data требует значительного объема памяти для хранения, что может быть дорого и сложно в реализации.

Дискретизация данных - некоторые алгоритмы ML, такие как дерево решений, могут работать только с дискретными данными. Для дискретизации данным в Big Data требуются нестандартные методы. Они трудоёмкие, затратные и непрактичные.

Стоимость - чтобы правильно настроить рабочее ПО необходимо много различных ресурсов. Поэтому использование ML не является практичным решением для всех.

Плюсы:

Автоматизация – ML сокращает время и рабочую нагрузку на человека, выполняя огромную часть аналитической работы без малейшего участия человека. Автоматизация надежна, эффективна и быстра.

Самосовершенствование – обрабатывая данные Big Data алгоритмы ML способны благодаря этим же данным учиться. При этом сложно определить до какого уровня ML сможет совершенствоваться.

Широкая сфера применения - эта технология имеет очень широкий спектр применения. ML играет роль практически во всех областях, таких как образование, медицина, наука, банковское дело и бизнес. Это создает больше возможностей.

Возможность предугадывать тренды – ML может анализировать большие объемы данных и обнаруживать определенные тенденции и закономерности, которые не были бы очевидны людям. Например, для веб-сайта электронной коммерции, такого как Amazon, он служит для понимания поведения пользователей в Интернете и истории покупок, чтобы помочь им найти правильные продукты, предложения и напоминания, относящиеся к ним.

Заключение. В статье обсуждается пересечение Big Data и ML, эти технологии содержат в себе большой потенциал для использования в широком спектре сфер. Big Data предоставляет объемы информации, которую можно использовать для обучения алгоритмов ML. Они же могут создавать закономерности и решения, которые люди не могут идентифицировать самостоятельно. В статье также рассматриваются преимущества и проблемы работы с Big Data и ML, в том числе потребность в мощных вычислительных ресурсах, проблемы с точностью данных и их источниками. В целом в статье подчеркивается огромное значение этих двух областей для преобразования различных отраслей, от здравоохранения до финансов.

Переход на ML может стать большим скачком для бизнеса. Это влечет за собой переопределение рабочих процессов, архитектуры, сбора и хранения данных, аналитики и других модулей.

Пошаговый подход лучше всего подходит для такого перехода. Во-первых, предприятиям необходимо разработать надежную стратегию на основе искусственного интеллекта и ML, которая соответствует их бизнес-целям. Во-вторых качественные данные являются ключом к реализации всего потенциала инструментов ML. Компании должны создать корпоративную культуру вокруг данных. Правильные люди и правильные данные могут иметь огромное значение. Наконец, время имеет решающее значение, и предприятия должны действовать быстро.

Поскольку объем данных со временем продолжает увеличиваться, сбор данных и управление ими становится важнейшей задачей для бизнеса. Управление и определение смысла собранных таким образом данных для улучшения маркетинговой стратегии и увеличения доходов – это большая битва. Таким образом, ML незаменимо для решения задач, связанных с Big Data, и раскрытия скрытых моделей, знаний и идей из наборов данных, чтобы превратить

потенциал последних в реальную ценность для принятия деловых решений и научных исследований. Все минусы использования этих технологий вопрос времени, так как ошибки ML при работе с Big Data говорят о том, что есть еще огромное пространство для совершенствования.

Список литературы

[1] Big Data + Machine Learning = Love// Хабрахабр. URL: https://habr.com/ru/company/first/blog/692978/ (дата обращения: 19.03.2023).

[2] Chithrai Mani. How Is Big Data Analytics Using Machine Learning?// Forbes Technology Council. URL: https://www.forbes.com/sites/forbestechcouncil/2020/10/20/how-is-big-data-analytics-using-machine-earning/?sh=602cf9aa71d2 (дата обращения: 20.03.2023).

[3] Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce// Хабрахабр. URL: https://habrahabr.ru/company/dca/blog/267361/ (дата обращения: 9.05.2019).

[4] N. Carah and D. Angus, "Algorithmic brand culture: participatory labour, machine learning and branding on social media," Media, Cult. Soc., vol. 40, no. 2, pp. 178–194, 2018

[5] N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective: Cambridge University Press, 2011.

BIG DATA AND MACHINE LEARNING

N.I. Potapenko
Senior Lecturer, Department of Engineering Psychology and Ergonomics

V.A. Bud-Husaim
Student of the Department of Engineering Psychology and Ergonomics

A.N. Vasilkova
Assistant of the Department of Engineering Psychology and Ergonomics, master

Department of Engineering Psychology and Ergonomics
Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus
E-mail: a.vasilkova@bsuir.by

Annotation. Big Data is the vehicle for solving and rethinking problems in IT, business, medicine, etc. The volume of data search continues to grow at an incredible rate, the ability to choose the right one has become an important key factor in identifying sources for studying algorithms. learn to neutralize mitochondrial data. The purpose of the article is to enable the use of representation as an organization, OD and big data in order to gain a competitive advantage in the important world, to manage the volume of data.

Keywords: big data, machine learning, algorithm, neural network, dataset processing, addiction, self-learning.