

Оглавление

Введение	5
Глава 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОБУЧЕНИЯ СТУДЕНТОВ BIG DATA, ОБЕСПЕЧИВАЮЩИЕ КОЛЛАБОРАЦИЮ МАТЕМАТИКИ И ЦИФРОВЫХ ТЕХНОЛОГИЙ.....	7
1.1. Проблемы использования Big Data.....	7
1.2. Анализ больших данных в облаках.....	13
1.3. Основные задачи машинного обучения. Обучение с учителем (supervised learning)	17
Глава 2. МЕТОДИКА ОБУЧЕНИЯ СТУДЕНТОВ BIG DATA, ОБЕСПЕЧИВАЮЩАЯ КОЛЛАБОРАЦИЮ МАТЕМАТИКИ И ЦИФРОВЫХ ТЕХНОЛОГИЙ.....	24
2.1. Построение прикладных задач, направленных на коллорабoрацию математики и цифровых технологий.....	24
2.2. Творческая задача. Построение трехфакторной модели при помощи Machine Learning	37
Заключение	64
Список использованной литературы	66

Введение

Реализация многих видов деятельности в современном обществе неразрывно связана с применением цифровых технологий, одним из ключевых моментов которой является коллаборация математики и цифровых технологий. Обучение Machine Learning начинается с построения математических моделей, которые позволяют показать универсальность математических уравнений и алгоритмов, дают возможность дифференцировать описания разнообразных по своей природе процессов.

В процессе изучения Machine Learning использование понятий, связанных непосредственно с математикой студентами экономического профиля аграрных вузов, позволяет избежать формального подхода к обучению, осуществлять межпредметные связи. Кроме того, у студентов формируется представление о роли математических методов в преобразующей деятельности, о соотношении реального и идеального, о характере отражения математикой явлений окружающего мира.

Информационные процессы искусственно порождаются людьми с помощью разнообразных технических устройств для осуществления различных действий с информацией и происходят в социотехнических системах, например, пилоты управляют различными видами транспорта на основе информации с бортовых приборов и технических систем. Появились беспилотные трамваи, поезда, автомобили. Результат фиксации, отображения информации на каком-либо материальном носителе, т. е. зарегистрированное на носителе представление сведений независимо от того, дошли ли эти сведения до какого-нибудь приемника и интересуют ли они его, называют данными.

Данные — это информация, представленная в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе.

Данные сами могут выступать как источник информации. Информация, извлекаемая из данных, может подвергаться обработке, и результаты обработки фиксируются в виде новых данных.

Большие данные или Big Data — это серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объемов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста данных, распределения по многочисленным узлам вычислительной сети, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

С одной стороны, термин Big Data относится просто к управлению и анализу больших объемов данных. С другой стороны, Big Data предполагают нечто большее, чем просто анализ огромных объемов информации. Проблема не в том, что организации создают огромные объемы данных, а в том, что бóльшая их часть представлена в формате, плохо соответствующем традиционному структурированному формату баз данных. Сегодня Big data — социально-экономический феномен, который связан с тем, что появились новые технологические возможности для анализа огромного количества данных. Big Data — это естественное развитие аппарата математической статистики и усовершенствованный метод обработки информации.

Под Big Data («большие данные») будем понимать технологии работы с информацией огромного объема и всевозможного состава, которая находится в разных источниках и часто обновляется с целью создания новых продуктов, увеличения эффективности работы, и повышения конкурентоспособности предприятия.

Глава 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОБУЧЕНИЯ СТУДЕНТОВ BIG DATA, ОБЕСПЕЧИВАЮЩИЕ КОЛЛАБОРАЦИЮ МАТЕМАТИКИ И ЦИФРОВЫХ ТЕХНОЛОГИЙ

1.1. Проблемы использования Big Data

Самая большая проблема больших данных — затраты на их обработку. Сюда входят расходы и на дорогостоящее оборудование (его приходится регулярно обновлять для поддержания минимальной работоспособности при увеличении объема данных), и на заработную плату квалифицированным специалистам, обслуживающим огромные массивы информации.

Вторая проблема связана с профессионализмом самого аналитика, поскольку ему необходимо обрабатывать большое количество информации. Если, например, исследование выдает не два-три, а большое количество результатов, очень сложно остаться объективным и выделить из общего потока данных только те, которые окажут реальное влияние на состояние какого-либо явления или процесса.

Следующая проблема — потеря информации. Меры предосторожности требуют не ограничиваться простым однократным резервированием данных, а делать хотя бы две-три резервные копии хранилища. С увеличением объема растут сложности с резервированием, и IT-специалистам требуется найти оптимальное решение данной проблемы.

Не менее важной проблемой является проблема конфиденциальности Big Data. При переходе большинства сервисов по обслуживанию клиентов в онлайн очень легко стать мишенью для киберпреступников. Даже простое хранение личной информации без совершения каких-либо интернет-транзакций может быть чревато нежелательными для клиентов облачных хранилищ последствиями.

Общие проблемы с производительностью, обеспечением надежности, оптимизацией хранения, вызванные новым форматом работы с данными, требуют постоянного совершенствования технологии Big Data.

К основным техникам и методам анализа и обработки данных можно отнести следующие.

Методы класса, или глубинный анализ (Data Mining).

Data Mining (добыча данных, интеллектуальный анализ данных, глубинный анализ данных) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Термин был введен математиком Григорием Пятецким-Шапиро в 1989 г.

Компания применяет Data Mining, когда у нее уже есть некий массив данных, который ранее был как-то обработан, а теперь он обрабатывается вновь, возможно, как-то иначе, чем прежде, для получения неких полезных выводов.

Data Mining решает следующие задачи: работа с данными (агрегация, анализ, описание), выявление взаимосвязей и построение трендов (возможно, с конечной целью предсказания).

Краудсорсинг. Позволяет получать данные одновременно из нескольких источников, причем количество последних практически не ограничено.

Краудсорсинг предлагает способ систематизации данных путем избавления от лишнего материала; категоризации и оценки нужной информации. Это позволяет получать доступ к ценной информации, содержащейся в недрах груды сырых данных. Систематизация большого объема данных методом краудсорсинга позволяет избежать значительных дополнительных расходов. Не случайно в список краудсорсинговых входят компании Yahoo, IBM, eBay и многие другие.

Смешение и интеграция данных (data fusion and integration). Позволяют интегрировать разнородные данные из разнообразных источников с целью проведения глубинного анализа (например, цифровая обработка сигналов, обработка естественного языка, включая тональный анализ, и др.).

Имитационное моделирование(simulation). Метод, позволяющий строить модели, описывающие процессы так, как они проходили бы в действительности. Имитационное

моделирование можно рассматривать как разновидность экспериментальных испытаний. Имитационное моделирование можно рассматривать как разновидность экспериментальных испытаний.

Пространственный анализ (spatial analysis). Класс методов, использующих топологическую, геометрическую и географическую информацию в данных.

Визуализация аналитических данных. Представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

А/В-тестирование. Данный метод предполагает выбор из всего объема данных контрольной совокупности элементов, которую поочередно сравнивают с другими подобными совокупностями, где был изменен один из элементов. Проведение подобных тестов помогает определить, колебания какого из параметров оказывают наибольшее влияние на контрольную совокупность. Благодаря объемам Big Data можно проводить огромное число итераций, с каждой из них приближаясь к максимально достоверному результату.

Прогнозная аналитика. Прогнозная аналитика действует множество методов из статистики, интеллектуального анализа данных, анализирует как текущие данные, так и данные за прошлые периоды, на основе которых и составляет прогнозы о будущих событиях. Модели прогнозирования выявляют связи среди многих факторов, чтобы сделать возможной оценку рисков или потенциала, связанного с конкретным набором условий. Итог использования прогнозной аналитики — принятие верных (максимально эффективных для бизнеса) решений [19].

С помощью моделей прогнозирования можно предсказать поведение потенциальных клиентов, выявить наиболее популярные продукты и услуги, понять, что движет клиентами, почему они уходят, и предотвратить это и т. д. Использование инструментов прогнозной аналитики помогает создать модель поведения клиентов, а значит, и увеличить прибыль компании.

Машинное обучение (Machine Learning). Предполагает эмпирический анализ информации и последующее построение алгоритмов самообучения систем. Машинное обучение — это метод анализа данных, основанный на построении автоматизированной аналитической модели. Используя математические алгоритмы анализа данных, машинное обучение позволяет находить скрытые факторы и зависимости, не будучи заранее запрограммированным на определенное место поиска. Машинное обучение способно адаптироваться и переобучаться на основе вновь поступивших данных для получения надежных и репрезентативных результатов. Итог машинного обучения — ценные предсказания, которые помогают принять лучшее решение и осуществить правильные действия в реальном времени без вмешательства человека.

По словам ученого в области аналитики Т. Дэвенпорта, в условиях быстро меняющихся, растущих объемов данных необходимо быстрое потоковое моделирование, чтобы не отставать, и это можно сделать с помощью машинного обучения. Люди могут создать одну или две хорошие модели в неделю, а Machine Learning — тысячи.

Машинное обучение продолжает развиваться и модернизироваться. В число наиболее важных разработок входят ансамблевые методы, в которых прогнозирование осуществляется на основе набора моделей, где каждая модель участвует в каждом из запросов, а также дальнейшее развитие нейронных сетей глубокого обучения, имеющих более трех слоев нейронов. Такие глубокие слои в сети способны обнаруживать и анализировать отображения сложных атрибутов (состоящие из нескольких взаимодействующих входных значений, обработанных более ранними слоями), которые позволяют сети изучать закономерности и обобщать их для всех входных данных. Благодаря своей способности исследовать сложные атрибуты сети глубокого обучения лучше других подходят для многомерных данных — именно они произвели переворот в таких областях, как машинное зрение и обработка естественного языка.

Сетевой анализ. Сетевой анализ — наиболее распространенный метод для исследования социальных сетей:

после получения статистических данных анализируются созданные в сетке узлы, т. е. взаимодействия между отдельными пользователями и их сообществами.

Использование Dark Data. Так называемые Dark Data (темные данные) — это вся неоцифрованная информация о компании, которая не играет ключевой роли при ее непосредственном использовании. Аналитики причисляют к темным данным информацию, которую сотрудники используют в работе только один раз. После этого она теряется на просторах неорганизованного контента. На практике это означает, что около 80 % документов в компании не используется повторно.

В управлении темными данными используют метаданные, или «данные о данных», для идентификации, связывания отдельных файлов, организации информации, отсылок на другие материалы. В совокупности это позволяет разблокировать темные данные и использовать их в работе.

С их помощью обеспечивается классификация данных по проекту, клиенту, рабочему процессу, статусу и другим факторам. Благодаря метаданным нетрудно проверить, правильно ли функционируют рабочие процессы, потоки документов и маршруты задач.

Метаданные выступают как микросигналы, которые добавляются к документам, и корпоративная информация, включая темные данные, становится доступной и экспортной.

Такой подход к управлению информацией позволяет включать темные данные в результаты поиска и распределять их по запросам пользователей. Организация получает полный обзор корпоративного контента.

Компания понимает, что не так важно, где хранится информация. Гораздо важнее, что собой представляет информация и как ее лучше классифицировать, чтобы получить максимальную пользу в ежедневной работе.

Искусственный интеллект. Искусственный интеллект (ИИ) как нельзя лучше подходит для обработки большого объема постоянно меняющейся информации. Машина делает все то же самое, что должен был бы сделать человек, но при этом вероятность ошибки значительно снижается (отрасль ИИ появилась еще в 1956 г.).

ИИ стимулирует идеи и ускоряет принятие решений. Более того, он сокращает трудоемкие ручные процедуры и ускоряет обслуживание внутренней инфраструктуры организации.

Blockchain. Blockchain — это технология распределенного реестра, которая позволяет ускорить и упростить многочисленные интернет-транзакции, в том числе международные. Благодаря этой технологии снижаются затраты на проведение транзакций.

Интеграция Blockchain с Big Data несет в себе синергетический эффект и открывает бизнесу широкий спектр новых возможностей, в том числе позволяя:

- получать доступ к детализированной информации о потребительских предпочтениях, на основе которых можно выстраивать подробные аналитические профили для конкретных поставщиков, товаров и компонентов продукта;
- интегрировать подробные данные о транзакциях и статистике потребления определенных групп товаров различными категориями пользователей;
- получать подробные аналитические данные о цепях поставок и потребления, контролировать потери продукции при транспортировке (например, потери веса вследствие усыхания и испарения некоторых видов товаров);
- противодействовать фальсификации продукции, отмыванию денег, мошенничеству и т. д.

Доступ к подробным данным об использовании и потреблении товаров в значительной мере раскрывает потенциал технологии Big Data для оптимизации ключевых бизнес-процессов, снижения рисков, появления новых возможностей создания продукции, отвечающей актуальным потребительским предпочтениям.

Технология распределенного реестра обеспечивает целостность информации, а также надежное и прозрачное хранение всей истории транзакций. Big Data, в свою очередь, предоставляет новые инструменты для эффективного анализа, прогнозирования, экономического моделирования и, соответственно, открывает новые возможности для принятия более взвешенных управленческих решений.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru