УДК 004.04

БОЛЬШИЕ ДАННЫЕ И МАШИННОЕ ОБУЧЕНИЕ

И.В. МИСЕВИЧ
(Представлено: канд. физ.-мат. наук, доц. Ю.Ф. ПАСТУХОВ)

Статья посвящена изучению Больших данных и их особенностей, актуальности применения и значение для нашей повседневной жизни. Особое внимание уделяется способам обработки такого вида данных, в частности, рассматривается машинное обучение как наиболее эффективный. На конкретных примерах показано то, как работают методы машинного обучения.

Актуальность данной работы обусловлена появлением нового типа данных – Больших данных, которые открывают новые возможности для практически каждой сферы общественной жизни. Проблема заключается в сложности обработки таких объёмов данных, для чего и было создано, большое количество устройств и программ, в частности, машинное обучение.

Цель – рассмотреть, что такое большие данные, машинное обучение и выяснить принцип работы машинного обучения на конкретном примере.

Задачи:

1. Поиск актуальной информации на заданную тему в различных письменных источниках и ресурсах Интернета;
2. Анализ найденной информации, её обобщение и систематизация, добавление собственных наблюдений;
3. Структурирование и изложение полученных данных;
4. Формулировка соответствующего вывода, ответ на поставленный в цели вопрос.

Методологическая основа данной работы включает в себя: анализ и обобщение специальной литературы, публикаций в периодических изданиях, наблюдение, описание, иллюстрирование примерами, формулировка собственного мнения и вывода.

У понятия «большие данные» нет однозначного определения, можно встретить множество трактовок и версий. Их объединяет одно — под большими данными подразумевается совокупность специальных технологий. Их используют для обработки значительно большего объёма данных (от петабайта: 1015 байт), чем это было до появления «больших данных». Данные определяют как множество объектов и множество соответствующих им ответов (откликов). Кроме того, большие данные должны работать с этими поступающими в большом количестве данными быстро, а также обрабатывать как структурированные, так и плохо структурированные данные.

Большие данные непрерывно накапливаются практически в любой сфере человеческой жизни. Это и социальные медиа, и медицина, и банковская сфера, и реклама, а также системы устройств, получающие многочисленные результаты ежедневных вычислений. Например, астрономические наблюдения и метеорологические сведения.

Информация с разнообразных систем слежения в режиме реального времени поступает на сервера компаний, использующих большие данные.

Технологии больших данных неотрывны от научно-исследовательской деятельности и коммерции. Более того, они начинают захватывать и сферу государственного управления – везде требуется внедрение все более эффективных систем хранения и детерминирование информации.

Существует большое количество техник и методов для анализа и обработки такой информации. Среди основных можно выделить следующие:

- **Методы класса или глубинный анализ (Data Mining).** Данные методы основаны на использовании особого математического инструментария в совокупности с достижениями из сферы информационных технологий.
- **Краудсорсинг.** Данная методика позволяет получать данные одновременно из практически неограниченного числа источников.
- **A/B-тестирование.** Из всего объема данных выбирается итоговая совокупность элементов, которую поочередно сопоставляют с другими похожими совокупностями, где был изменён один из элементов, что помогает определить, изменения какого из параметров оказывают наибольшее влияние на совокупность.
- **Прогнозная аналитика.** Данный метод направлен на предугадывание и планирование того, как будет вести себя подконтрольный объект, чтобы принять наиболее выгодное в этой ситуации решение.
- **Машинное обучение (искусственный интеллект).** Метод основан на эмпирическом анализе информации и последующем построении алгоритмов самообучения систем.

- **Сетевой анализ.** После получения статистических данных анализируются созданные в сетке узлы, то есть взаимодействия между отдельными пользователями и их сообществами.

Как уже было отмечено выше, машинное обучение является одним из методов обработки больших данных. Рассмотрим его подробнее.

Машинное обучение – это математическая дисциплина, в рамках которой решается задача поиска закономерностей в эмпирических данных; на их основе алгоритм может давать определенные прогнозы. Машинное обучение можно отнести к методам искусственного интеллекта, так как оно не решает задачу напрямую, а обучается применять решение для множества похожих задач.

Машинное обучение находится между математической статистикой, методов оптимизации и классических математических дисциплин, но имеет также и отличительную особенность, связанную с проблемами эффективности вычислений и переобучения. Многие методы также тесно связаны с извлечением информации и интеллектуальным анализом данных (Data Mining).

Машинное обучение избавляет программиста от необходимости подробно объяснять компьютеру, как именно решать проблему. Вместо этого компьютер обучают самостоятельно находить решение.

Алгоритм получает набор необходимых данных, а затем использует их для обработки запросов. К примеру, вы можете загрузить в машину код нескольких фотографий со следующим описанием: «на этом фото изображена собака» и «на этом фото нет собаки». Если после этого загрузить в компьютер большое число новых изображений, он начнёт самостоятельно сортировать снимки.

Машины учатся видеть изображения и классифицировать их, как в примере с фото. Они могут распознавать текст, числа, людей и местность на этих изображениях. Компьютеры не просто выявляют отличительные особенности для сортировки, но и учитывают контекст их употребления.

Конечно, не стоит ожидать 100% правильного результата, случаются и ошибки. Верные и ошибочные результаты распознавания попадают в базу данных, тем самым давая возможность программе учиться на своих ошибках и лучше справляется с поставленной задачей. Теоретически, процесс совершенствования может развиваться бесконечно долго. В этом и состоит суть процесса обучения.

Существует несколько общих методов машинного обучения: обучение с учителем (самый распространённый и рабочий), обучение без учителя и обучение с подкреплением.

Концепция первого метода состоит в том, что в систему загружается тренировочный набор данных – «обучающая выборка», в которой информация разбита на пары: данные ввода и данные вывода. Задача компьютера – понять логику, связывающую пары, создать алгоритм и с его помощью объединять в пары новые данные. Система постоянно совершенствуется, проанализированные данные становятся ее «опытом», она учитывает ошибки и стремится минимизировать их допущение в будущем. Так происходит обучение.

Пример: имеются данные о 1 000 000 квартир в Москве; о каждой известна площадь, количество комнат, этаж, месторасположение, наличие парковки, и так далее. Кроме того, известна стоимость каждой квартиры. Задача - построение модели, которая на основе данных признаков будет определять стоимость квартиры. Это классический пример обучения с учителем, где у нас есть данные (различные параметры для каждой квартиры) и отклики (стоимость квартиры). Такая задача называется задачей регрессии.

Обучение без учителя – тренировка без подсказок. Обучающая выборка состоит только из данных ввода. Задача программы – выявить всевозможные зависимости и связи между заданными параметрами. Так как данные не парны, у системы нет «шпаргалки» с правильными ответами. Выводы о наличии связей система делает самостоятельно. Ожидаемый результат – разделение информации на кластеры или обнаружение отклонений от введённых параметров. С каждым новым процессом система будет учиться более точно классифицировать данные.

Пример: допустим нам известны данные о росте и весе некоторого числа новорождённых. Необходимо сгруппировать данные на 3 группы, чтобы для каждой выпустить детские ползунки соответствующего размера. Это задача кластеризации. Нужно отметить, что разделение на кластеры является не таким явным и часто нет «правильного» разделения.

Ещё один метод – обучение с подкреплением. Оно подразумевает взаимодействие системы со средой. Такое взаимодействие вызывает отклик, положительный или отрицательный. Это помогает программе постепенно обнаружить оптимальные пути для стабилизации отклика.

После тысяч часов вычислений и операций, обучаемая любым из методов система готова к неизвестности. Ее чётко отработанные алгоритмы способны на прогнозирование, классификацию, кластеризацию принципиально новых данных. В процессе обработки модель продолжит учиться и совершенствоваться. Процесс обучения идет до тех пор, пока пополняется база.

Методы машинного обучения позволяют лучше понять клиента, облегчить поиск товаров, повысить конверсию, оценить риски, связанные с теми или иными инвестициями. Согласно опросу, проведённому MIT Review Custom и Google Cloud, 60% опрошенных, представляющих самые разнообразные компании, заявили, что машинное обучение для них является основным способом обработки массивов данных. Среди основных мотивов участники опроса называли стремление извлечь новые знания из своих данных, приобрести конкурентное преимущество, ускорить анализ информации и выпуск продукции нового поколения.

Два года назад Facebook, алгоритмы которого построены с помощью машинного обучения, открыл код используемого им программного обеспечения. В прошлом году открытой стала библиотека ПО для машинного обучения Google'a TensorFlow. Она помогает IT-специалистам понять, как работают модели машинного обучения, и встраивать их в свои продукты. Twitter, Uber.

Эти платформы значительно снизили важность понимания алгоритмов, на которых основывается машинное обучение. С помощью открытых инструментов и коммерческих облачных решений использование машинного обучения стало более доступно. Эксперты прогнозируют, что очень скоро любой обыватель вместо нагромождений кода и алгебраических выкладок, сможет применять механизмы машинного обучения, пользуясь понятным интерфейсом.

Таким образом, большие данные становятся неотъемлемой частью нашего повседневного существования. Каждый день огромное количество информации о нас и о наших предпочтениях помогает создавать сложные концепции умных домов, умных городов, интернета вещей и так далее. Тем самым, анализ больших данных помогает улучшить и преобразовать почти все стороны нашей жизни.

Big Data открывает перед нами новые горизонты в планировании производства, образовании, здравоохранении и других отраслях. Если и дальше будет продолжаться их развитие, то технологии больших данных смогут поднять информацию, как фактор производства, на совершенно новый качественный уровень.

ЛИТЕРАТУРА

1. Воронцов, К. В. Лекции по машинному обучению. www.MachineLearning.ru. 2004-2016.
2. Паньшин И. Машинное обучение [Электронный ресурс] – Режим доступа: https://newtonew.com/tech/machine-learning-novice. Дата доступа: 25.09.2020.
3. Соколов Е. Введение в машинное обучение и анализ данных [Электронный ресурс] – Режим доступа: http://docplayer.ru/73701440-Vvedenie-v-mashinnoe-obuchenie-i-analiz-dannyh.html. Дата доступа: 25.09.2020.
4. Машинное обучение: искусственный интеллект помогает упорядочить хаос больших данных [Электронный ресурс] – Режим доступа: http://sap-technology.rbc.ru/mashinnoe-obuchenie.html. Дата доступа: 25.09.2020.
5. Технологии Big Data: как использовать большие данные в маркетинге [Электронный ресурс] – Режим доступа: https://www.uplab.ru/blog/big-data-technologies/. Дата доступа: 25.09.2020.
6. Большие данные: новая теория и практика [Электронный ресурс] – Режим доступа: https://www.osp.ru/os/2011/10/13010990/. Дата доступа: 25.09.2020.
7. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce [Электронный ресурс] – Режим доступа: https://habrahabr.ru/company/dca/blog/267361/. Дата доступа: 25.09.2020.
8. Веретенников А. В. BigData: анализ больших данных сегодня // Молодой ученый. – 2017. – №32. – С. 9-12 [Электронный ресурс] – Режим доступа: https://moluch.ru/archive/166/45354/. Дата доступа: 25.09.2020.