УДК 004.75

Методы анализа больших данных

Коновалов Эльдар Наилевич, студент направления
«Информационные системы и технологии»;
Штырова Ирина Анатольевна, кандидат технических наук, доцент кафедры
«Информационные системы и технологии»
Балаковский инженерно-технологический институт – филиал федерального
государственного автономного образовательного учреждения высшего образования
«Национальный исследовательский ядерный университет «МИФИ», г. Балаково

В данной статье рассмотрены основные понятия и области применения больших данных, приведены примеры их источников. Также рассмотрен ряд важных характеристик, сформулированы основные принципы работы с большими данными. Рассмотрены наиболее распространенные методы анализа больших данных, такие как глубинный анализ данных, краудсорсинг, машинное обучение, искусственные нейронные сети. Методы работы с большими данными позволяют выявлять ранее неизвестную информацию, а также различные закономерности, что позволяет принимать наиболее эффективные решения.

С развитием информационных технологий и техники постоянно растут объемы информации, генерируемые человечеством. Каждые 40 месяцев на протяжении 80-х годов 20 века удваивалось общее количество информации, которые могут хранить все технические средства мира, а в настоящий момент ежедневная генерация исчисляется в 2,5 экзабайта. Стандартными методами обработки информации с целью ее эффективного использования являются методы математической статистики, теории баз данных [1]. Однако все в большей степени хранящиеся данные являются неструктурированными и изменчивыми. Источниками таких данных могут выступать постоянно получаемые данные с различных измерительных устройств, устройств аудио- и видео-регистрации, потоки сообщений в социальных сетях, устройства класса «интернет-вещей», а также информация, находящаяся в обращении внутри предприятий и организаций. Необходимость выявления закономерностей в больших массивах неструктурированных данных обусловило появление новых методов обработки и анализа больших данных.

Под большими данными (Big Data) понимают массивы данных, которые нельзя обработать, проанализировать при помощи обычных аналитических методов в связи с очень большим объемом и сложной структурой данных [2]. Сложность анализа при работе с большими данными обуславливается особенностями получения, разделения,

хранения, передачи, процесса визуализации, сохранения анонимности полученных данных. Анализ больших данных, как правило, предполагает применение определенных методов работы с данными для того, чтобы выявить различные закономерности, а также другую полезную информацию. Точность при работе над большими данными позволяет получать информацию с наиболее высокой степенью полезности и принимать лучшие решения с целью повышения эффективности работы предприятия или организации.

Примером области применения больших данных может являться сельское хозяйство. Сенсоры, датчики и другие технические устройства устанавливаются на территории фермерских хозяйств и объединяются в единую сеть, основанную на больших данных. Подобные сети позволяют фермерам получать визуализированное представление всех процессов сельскохозяйственного производства, а также интегрировать в систему данные о потребностях рынка. Фермерам доступна информация о состоянии почвенных ресурсов, погоде и влажности, степени созревания плодов. Это позволяет в режиме реального времени оптимизировать хозяйство под текущие условия рынка [3].

Другим примером может являться сфера онлайн-торговли. Компании используют большие данные для анализа интернет-покупок, они сравнивают покупки, которые совершают их клиенты, с тем, что покупают другие пользователи. Результатом такого анализа является информация о том, что может понадобиться их клиентам сейчас или в будущем.

Большие данные также используются в медицине. При помощи технологий происходит сбор, обмен и анализ накопленных данных врачами и учеными, что позволяет совершенствовать медицинскую диагностику, лечение пациентов. Анализ аккумулируемых статистических данных позволяет предупреждать риски инфекций и заболеваний [4]. Также большие данные помогают создавать умную систему карантина в периоды эпидемий. Ряд стран во время эпидемии коронавируса активно применяет приложения, использующие возможности больших данных, для анализа социальных контактов пользователя и его перемещения, что позволяет в случае его заболевания свести опасные последствия к минимуму.

Для анализа больших данных кроме физического объема существует ряд других важных характеристик, которые определяют сложность анализа и обработки данных. Изначально существовал набор признаков VVV (volume, velocity, variety – объем данных, скорость приращения данных и необходимость их быстрой обработки, возможность одновременной обработки данных). Этот набор признаков был выработан

в 2001 году компанией Meta Group для того, чтобы показать равнозначность всех трех аспектов при управлении данными [5].

Потом появились и расширенные версии изначального набора признаков, такие как VVVV (была дополнительно включена в набор veracity – достоверность), VVVVV (были дополнительно включены в набор viability – жизнеспособность и value – ценность), VVVVVVV (были дополнительно включены в набор variability – переменчивость и visualization – визуализация).

Следуя вышеописанным признакам, для больших данных были сформулированы следующие принципы работы.

Горизонтальная масштабируемость – это основной принцип обработки для больших данных. Больших данных становится постоянно все больше и больше с каждым днем. Поэтому нужно постоянно увеличивать число вычислительных узлов, по которым будут распределены необходимые данные, с учетом того, что обработка данных не должна приводить к падению производительности.

Отказоустойчивость. Этот принцип является прямым следствием горизонтальной масштабируемости. В связи с тем, что вычислительных узлов может быть очень много и их количество будет увеличиваться, не исключено, что будет расти вероятность сбоев и выхода из строя оборудования. Такая возможность должна обязательно учитываться при работе с большими данными и предполагать введение превентивных мер.

Локальность данных. Издержки на передачу данных могут сильно вырастать, так как, в связи с высокой распределенностью, данные зачастую находятся на одном сервере, а обрабатываются на другом. Разумно проводить обработку данных на той же машине, где эти данные и хранятся.

Вышеперечисленные принципы сильно отличаются от тех, которые характерны для общераспространенных, централизованных и вертикальных моделей хранения с высокой степенью структурированности данных. Поэтому для работы с большими данными постоянно разрабатывают новые методы.

Согласно определениям международной консалтинговой компании McKinsey, выделяют следующие методы работы с большими данными.

Data Mining (глубинный анализ данных) – это комплекс методов, задача которых – обнаружить ранее неизвестные закономерности и полезную информацию в некотором массиве данных. К методам этого класса относят обучение ассоциативным правилам, категорирование данных, регрессионный анализ и анализ кластеров [6].

Смешение и интеграция данных – это набор техник, предназначенных для интегрирования разнородных данных для их последующего глубокого анализа. Разнородные данные из разных источников приводятся к единому формату, если имеется несколько источников данных об одном объекте, то данные дополняются с целью получения полного представления, отсеиваются избыточные данные.

Машинное обучение – набор методов искусственного интеллекта, ключевой особенностью которых является решение задач за счет обучения в процессе многократного решения сходных задач. Используются методы обучения без учителя – это методы, относящиеся к задачам кластеризации, к задачам поиска закономерностей, частотных правил и т. д.; методы обучения с учителем, которые позволяют выполнять прогнозирование; методы классификации [7].

Искусственные нейронные сети – это программный комплекс и математическая модель, спроектированные по подобию биологических нейронных сетей, конечной задачей сетей является решение задач оптимизации и моделирования [8]. С точки зрения машинного обучения, нейронная сеть представляет собой частный случай методов распознавания образов, дискриминантного анализа, методов кластеризации и т. п. С математической точки зрения, обучение нейронных сетей – это многопараметрическая задача нелинейной оптимизации. Наиболее популярные нейронные сети – персептроны (Perceptrons, P), сети прямого распространения (Feed Forward, FF), рекуррентные нейронные сети (RNN), сети Кохонена (KohonenNetwork, KN), сети Хопфилда (Hopfield Network, NN).

Имитационное моделирование – метод построения моделей, задача которых описывать необходимые процессы, как если бы они происходили в реальности. Имитационное моделирование позволяет выявить структуру исследуемых данных, динамику изменений, способствуя быстрому реагированию на изменение разнородных признаков.

Пространственный анализ – это набор методов, позволяющих извлекать полезную информацию топологического и геометрического типа из данных. Данный метод используется для решения сложных локационно-ориентированных задач, в которых имеется необходимость находить закономерности, оценивать тенденции.

Статистический анализ – методы математической статистики, такие как:

– корреляционный анализ, позволяющий выявить взаимосвязи и оценить влияние изменения одних данных на другие;

– анализ временных рядов, который позволяет оценить интенсивность и частоту изменений данных с течением времени;

– A/B-тестирование, используя этот метод сравнивают наборы тестовых и контрольных групп, в которых некоторые значения показателей намеренно изменены, для того чтобы узнать, какое из изменений приносит наилучший эффект.

Визуализация аналитических данных – графическое представление информации в виде гистограмм, диаграмм, рисунков, а также применяя интерактивные возможности и анимацию. Может применяться как для формирования исходных данных для дальнейшего анализа, так и для получения конечных результатов. Метод позволяет наглядно представить самые важные моменты анализа.

Подводя итог можно с уверенностью сказать, что большие данные прочно вошли в инструментарий сегодняшнего дня. Использование больших данных позволяет выявлять ранее неизвестные закономерности и использовать их на благо человека.

Литература

1. Фролов, М. В. Использование OLAP-технологий для оптимизации обработки данных в информационной системе вузовского центра дополнительного образования / М. В. Фролов, О. В. Виштак // Сборник трудов III Всероссийской научно-практической конференции «Современные технологии в атомной энергетике». – М.: НИЯУ МИФИ; Балаково: БИТИ НИЯУ МИФИ, 2017. – Т. 1. – С. 114-116.

2. UPLAB: [сайт]. – URL: https://www.uplab.ru/blog/big-data-technologies (дата обращения: 06.04.2021). – Текст: электронный.

3. Половинченко, М. И. Большие данные и их применение в агробизнесе / М. И. Половинченко, В. С. Елисеев // Мехатроника, автоматика и робототехника. – 2021. – № 7. – С. 46-49.

4. Ильяшенко, В. М. Применение технологии больших данных в здравоохранении / В. М. Ильяшенко // Сборник научных статей по итогам работы круглого стола с международным участием «Мир в эпоху глобализации экономики и правовой сферы: роль биотехнологий и цифровых технологий». – М.: Учебно-курсовой комбинат «Актуальные знания», Ассоциация «Союз образовательных учреждений», 2021. – С. 247-249.

5. IT.Enterprise: [сайт]. – URL: https://www.it.ua/ru/knowledge-base/technology-innovation/big-data-bolshie-dannye (дата обращения: 05.04.2021). – Текст: электронный.

6. Замятин, А. В. Интеллектуальный анализ данных: учебное пособие / А. В. Замятин. – Томск: Издательский Дом Томского государственного университета, 2016. – 120 с.

7. Рябенков, Р. А. Машинное обучение / Р. А. Рябенков, И. В. Михеев // Сборник трудов II Международной научно-практической конференции «Современные технологии и автоматизация в технике, управлении и образовании». – М.: НИЯУ МИФИ; Балаково: БИТИ НИЯУ МИФИ, 2020. – С. 224-228.

8. Сидоренко, Д. Е. Анализ возможности использования нейронных сетей для оценки качества программных продуктов / Д. Е. Сидоренко, И. В. Михеев // Сборник трудов II Международной научно-практической конференции «Современные технологии и автоматизация в технике, управлении и образовании». – М.: НИЯУ МИФИ; Балаково: БИТИ НИЯУ МИФИ, 2020. – С. 242-247.

УДК 373.1

Использование информационных технологий в инклюзивном образовании

[1]Куликова Елена Геннадьевна, учитель;

[2]Штырова Ирина Анатольевна, кандидат технических наук, доцент кафедры «Информационные системы и технологии»;

[1]Муниципальное автономное общеобразовательное учреждение «Основная общеобразовательная школа села Быков Отрог» Балаковского района Саратовской области;

[2]Балаковский инженерно-технологический институт – филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ», г. Балаково

Использование информационных технологий в образовательном процессе играет важную роль для коррекции и компенсации недостатков развития детей, относящихся к категории учащихся с ограниченными возможностями здоровья, позволяет вовлечь их в процесс обучения, создать условия для развития познавательной деятельности. Это является основой для создания адаптивной образовательной среды, обеспечивающей индивидуализацию обучения с учетом особенностей и образовательных потребностей конкретного обучающегося.

Инклюзия в образовании признана Россией на государственном уровне. Под инклюзивным образованием понимается целенаправленный процесс обучения и социализации лиц с особыми образовательными потребностями в образовательных организациях общего, профессионального и дополнительного образования. Практическая реализация инклюзивного образования остается очень сложной проблемой. Образовательный процесс исключительных детей – это целый комплекс