УДК 04 **Информационные технологии**

Матвеев Андрей Николаевич, *студент, каф. МОЭВМ,* *Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург*

e-mail: andrewsamsunguser@gmail.com

Семёнов Александр Николаевич, *студент, каф. МОЭВМ,* *Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург*

e-mail: sashaux@yandex.ru

ИСПОЛЬЗОВАНИЕ DATA SCIENCE И МАШИННОГО ОБУЧЕНИЯ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ

Аннотация: В статье анализируется использование Data Science и машинного обучения в интеллектуальных системах. Рассматриваются характеристика и типология больших данных и основные проблемы, возникающие при работе с ними. Выявляются возможности и алгоритм работы Data Science в интеллектуальных системах. Приводятся общая задача, модели и основные методы машинного обучения, которые используются в интеллектуальных системах для предсказания событий и улучшения понимания анализируемой информации.

Ключевые слова: большие данные, Data Science, аналитика, машинное обучение, интеллектуальные системы.

Annotation: The article analyzes the use of Data Science and machine learning in intelligent systems. The characteristics and typology of big data and the main problems that arise when working with them are considered. The possibilities and algorithm of Data Science work in intelligent systems are revealed. The general task, models and main methods of machine learning that are used in intelligent systems to

predict events and improve the understanding of the analyzed information are given.

Keywords: big data, data science, analytics, machine learning, intelligent systems.

Одно из актуальных направлений российской государственной политики, соответствующее мировым тенденциям – освоение цифрового пространства путём ускоренного внедрения цифровых технологий в экономическую и социальную сферы [1]. Массовая цифровизация привела к накоплению и непрерывной генерации огромных массивов цифровой информации – больших данных, которые не поддаются обработке традиционными методами [2]. Обработка и анализ больших данных осуществляется при помощи технологий Data Science, которые позволяют выявлять в данных неочевидные и ценные для принятия решений связи и закономерности. Результаты аналитики используются интеллектуальными системами, которые позволяют находить оптимальные технологические, производственные и управленческие решения и повышать эффективность деятельности вне зависимости от её направления. Перспективность применения технологий Data Science и машинного обучения в интеллектуальных системах обуславливает актуальность исследования их специфики.

Целью работы является изучение использования Data Science и машинного обучения в интеллектуальных системах. Для её достижения были использованы аналитический, синтетический, индуктивный и дедуктивный методы обработки тематических исследований, научных публикаций и релевантных литературных источников.

Большие данные характеризуются четырьмя категориями [3]:

- разнообразие – наличие структурированных, полуструктурированных и неструктурированных данных, собранных людьми и машинами;
- скорость – оперативность сбора и обработки данных для получения ожидаемых результатов;
- объём – большие объёмы данных, постоянно поступающих из разных

источников;

- достоверность – наличие в данных погрешностей, аномалий и шумов.

Структурированные данные, составляющие порядка 20 % от всех существующих данных, обычно хранятся в организованных в соответствии с различными моделями базах данных и используются преимущественно в программировании. Полуструктурированные данные не соответствуют формальной модельной структуре, однако содержат маркеры или теги для идентификации информации, что облегчает процесс её обработки. Неструктурированные данные не имеют организации и не соответствуют моделям данных, хотя именно на них приходится большая часть информации, генерируемой машинами и человеком.

Можно выделить следующие проблемы работы с большими данными [4]:

- большой объём данных, который требует применения дорогостоящих технологий хранения и обработки;
- сложность хранения данных с поддержанием их целостности, доступности и конфиденциальность;
- неструктурированный вид данных, бессистемно объединяющий разноформатную информацию;
- сложность структурирования, сортировки и распределения данных при формировании выборок и поиске отдельных элементов;
- низкая скорость обработки, приводящая к продолжительному ожиданию ответа при поиске определённых позиций и устареванию данных в процессе обработки;
- отсутствие эффективных алгоритмов обработки данных, которые учитывали бы объём хранилища, структуру и методы поиска элементов;
- большое количество шумов и помех, которые необходимо учитывать при работе с наборами данных.

Для решения указанных проблем и извлечения пользы из массивов неструктурированных данных используется концепция Data Science, объединяющая методы математической статистики и машинного обучения [5].

Основная задача Data Science – выявить в данных нетривиальные, ранее неизвестные и практически ценные связи и закономерности и создать на их основе прогнозную модель процесса, которая позволит найти оптимальное решение поставленной задачи.

Data Science является мультидисциплинарной областью, поэтому включает множество методов и алгоритмов из разных областей: прикладной статистики, эффективных вычислений, теории баз банных, оперативной аналитической обработки, экспертных систем, распознавания образов, визуализации данных и нейронных сетей. В интеллектуальных системах Data Science работает по следующему алгоритму [6]:

- сбор целевого набора данных из совокупности исходных сведений;
- предварительная обработка данных – очищение данных для определения стратегий обработки и учёта информации о временной последовательности сведений;
- преобразование данных – уменьшение и проецирование данных при помощи различных методов преобразования;
- интеллектуальный анализ данных – извлечение шаблонов из данных;
- интерпретация или оценка данных – извлечение знаний из выявленных шаблонов.

За интерпретацию выделенных данных и нахождение оптимальных решений в интеллектуальных системах отвечают технологии машинного обучения – индуктивные алгоритмы на основе статистических методов и нейронных сетей [7]. Общая задача машинного обучения заключается в нахождении отображения функции $f:X \to Y$ по подмножествам заведомо известных пар ($x \in X_l$, $y \in Y_l$), где X – множество наблюдаемых событий или объектов, Y – множество исходов, которые позволяют системе выполнять поставленные задачи, множество $X_l \subset X$ – события или объекты с известными исходами $Y_l \subset Y$. Пары (x, y) являются прецедентами, множества X_l и Y_l, выступающие исходными данными для нахождения целевой функции f, – обучающей выборкой. На практике задача машинного обучения ограничивается

отсутствием строгой корреляции наблюдаемых событий с исходами, гипотезой о решающей функции f', параметрами событий и исходов и требованием одновременного соблюдения точности, обобщённости и простоты искомого отображения.

В основе работы технологий машинного обучения лежит процесс обучения алгоритма на специально подобранном наборе данных [8]. Обучение может осуществляться по одной из трёх моделей:

1. Обучение с учителем. В этой модели данные представляют собой множества объектов и ответов, зависимость между которыми неизвестна. Конечная совокупность «объект-ответ» образует обучающую выборку, на основе которой строится алгоритм, способный находить зависимости и давать точный ответ для нового набора объектов без ответов.

2. Обучение без учителя. В этой модели известны только описания множества объектов, то есть обучающая выборка не содержит ответов. Алгоритму необходимо находить зависимости, взаимосвязи и закономерности между объектами.

3. Обучение с подкреплением. Этот тип обучения предполагает обучение на практике, для которого используются агенты обучения, обеспечивающие активное принятие решений и обучение на собственных результатах.

Основными методами обработки данных и машинного обучения в интеллектуальных системах являются [9-11]:

1. Задачи классификации. Подразумевают разделение объектов на классы, множество которых задано, по схожим признакам на основе определённых закономерностей.

2. Задачи регрессии. Позволяют получить численное значение неизвестного параметра с использованием известных параметров исследуемого объекта. При этом число классов заранее не определяется.

3. Задачи ассоциации. Их цель заключается в обнаружении частых ассоциаций между объектами или событиями. Получаемые закономерности представляются в виде совокупности правил, которые можно использовать для

предсказания событий и улучшения понимания анализируемой информации.

4. Задачи кластеризации. В их основе лежит выявление во всём множестве данных независимых групп и их параметров. Задачи позволяют сгруппировать однородные объекты, сократить их количество и облегчить последующий анализ.

5. Задачи ранжирования. Используются в системах информационного поиска и рекомендательных системах. Их цель – определить для каждого объекта приоритет в выдаче как результат некоторого запроса.

Указанные задачи делятся на описательные и предсказательные. Первые помогают повысить качество понимания анализируемых данных, вторые – построить модель по полученному результату и предсказать течение процесса или поведение системы.

Таким образом, стремительное развитие машинного обучения, основанного на Data Science, обусловлено его высокой эффективностью и универсальностью: поскольку любые процессы связаны с данными, машинное обучение может быть адаптировано для улучшения любого типа процессов. Использование интеллектуальных систем позволяет оптимизировать организационные системы и технологические процессы, благодаря чему повышается конкурентоспособность не только отдельных предприятий, но и экономики в целом.

Библиографический список:

1. Указ Президента РФ от 07.05.2018 N 204 (ред. от 21.07.2020) «О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года» [Электронный ресурс] // Сайт Президента России. – URL: http://kremlin.ru/events/president/news/57425 (дата обращения: 08.06.2023).

2. Стеценко А.П. Инновации: Data Science // Успехи в химии и химической технологии. – 2021. – Т. 35, № 1 (236). – С. 62-64.

3. Возникающие тенденции в области технологий 2021: искусственный интеллект и большие данные для развития 4.0 / Станкович М., Гарба А.А.,

Нефтенов Н. – Женева: Международный союз электросвязи, 2021. – 108 с.

4. Менщиков А.А. Основные проблемы использования больших данных в современных информационных системах / А.А. Менщиков, В.Э. Перфильев, М.Ю. Федосенко, И.Р. Фабзиев // Столыпинский вестник. – 2022. – Т. 4, № 1. – С. 316-329.

5. Часовских В.П. Области искусственного интеллекта: лекция курса «Системы искусственного интеллекта». – Екатеринбург: УГЭУ, 2022. – 26 с.

6. Клюева И.А. Современные возможности и примеры внедрения машинного обучения // Оригинальные исследования. – 2021. – Т. 11, № 7. – С. 12-32.

7. Михайлюк А., Петренко П. Машинное обучение и онтологии как два подхода к построению интеллектуальных систем // Information Technologies & Knowledge. – 2019. – Т. 13, № 1. – С. 55-75.

8. Кондратёнок Е.В., Макареня С.Н. Машинное обучение как инструмент анализа данных // Информационные технологии в образовании, науке и производстве: IX Международная научно-техническая интернет-конференция / сост. Е.А. Хвитько. – Минск: БНТУ, 2022. – С. 304-309.

9. Савенков П.А. Использование методов и алгоритмов машинного обучения в системах поддержки принятия управленческих решений // Вестник науки и образования. – 2019. – № 1-2 (55). – С. 23-25.

10. Кадиев Ш.К. Обзор исследований в области классификации для машинного обучения при разработке интеллектуальных систем поддержки принятия управленческих решений / Ш.К. Кадиев, Р.Ш. Хабибулин, П.П. Годлевский, В.Л. Семиков // Технологии техносферной безопасности. – 2020. – № 3 (89). – С. 20-29. DOI: 10.25257/TTS.2020.3.89.20-29.

11. Миронов А.М. Машинное обучение. Часть 1: учебник для вузов. – М.: МАКС Пресс, 2018. – 90 с.