«Сейфуллин оқулары – 16: Жаңа формациядағы жастар ғылыми – Қазақстанның болашағы» атты халықаралық ғылыми-теориялық конференциясының материалдары = Материалы Международной научно-теоретической конференции «Сейфуллинские чтения – 16: Молодежная наука, новой формации – будущее Казахстана». - 2020. - Т.I, Ч.3 - С.195-197

АНАЛИЗ ДАННЫХ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ПРОГНОЗА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Тулегенов Т.Н.

Интенсивное развитие технологий к настоящему времени привело к экспоненциальному росту собранных данных: 90% всех цифровых данных были созданы за последние два года [1]. К 2025 году количество одних только устройств т.н. «Интернета вещей», имеющих доступ к глобальной сети, составляет порядка 38 млрд [2]. Возрастающий тренд на увеличение объема структурированных и неструктурированных данных актуализировал стремительное развитие технологий на основе анализа больших данных. При этом остро встала проблема извлечения полезной информации из данных самой разной природы с учетом обеспечения вычислительной эффективности обработки.

Изучение состояния исследований и разработок по указанной проблеме показывает, что существующие методы и алгоритмы обработки больших данных в отдельных случаях не в полной мере отвечают потребностям заказчиков из самых разных отраслей деятельности.

Указанное обуславливает наличие в составе мировых лидеров ИТ-отрасли, таких как компании Google, Facebook, Amazon, Microsoft, IBM подразделений, занимающихся изучением перспектив применения технологий машинного обучения для обработки больших данных и созданием продуктов на их основе [3].

В Республике Казахстан в соответствии с государственной программой «Цифровой Казахстан» на 2018-2022 годы одним из ключевых направлений определен переход на цифровое государство [4], которое окажет значимый эффект с точки зрения экономии бюджетных средств и мультипликативного роста ВВП. В частности, по данному направлению реализуется аналитическая платформа Smart Data Ukimet на основе технологий машинного обучения. К текущему моменту уже реализован ряд кейсов: проведена работа по трудоустройству населения, выделению бесплатных лекарственных средств, выделению грантов по программе «Болашак».

Таким образом, в контексте цифровизации всех областей жизнедеятельности, промышленности и науки, а также стремления Республики Казахстан к выходу из зоны «инновационной пассивности», существующая потребность в создании и использовании систем поддержки решений на основе технологий машинного обучения во всех отраслях экономики будет только увеличиваться.

Вместе с тем, использование технологий машинного обучения является лишь одним из этапов в процессе анализа данных используемых в системах поддержки принятия решений и обработки информации. Одной из распространенных является следующая условная группировка основных этапов процесса анализа данных [5]:

1. Постановка задачи анализа;
2. Предварительная подготовка данных для последующего анализа;
3. Применение методов машинного обучения и построение моделей;
4. Реализация построенных моделей и оценка их эффективности;
5. Интерпретация результатов экспертом.

Целью работы является разработка модели, позволяющей в зависимости от заданного критерия эффективно выявлять и описывать скрытые зависимости между признаками при помощи различных функций, её реализация в форме программного прототипа, а также экспериментальная оценка по таким характеристикам как масштабируемость, вычислительная эффективность и точность в задачах принятия решений [6].

В соответствии с поставленной целью в работе будут решены следующие задачи исследования:

1. Исследование основных методов отбора признаков данных для задач машинного обучения.
2. Проведение анализа данных, используемых для решения задач машинного обучения.
3. Разработка структурно-функциональной модели, реализующей математически корректный, масштабируемый и вычислительно эффективный алгоритм определения скрытых зависимостей между признаками данных.
4. Программная реализация разработанной структурно-функциональной модели.
5. Оценка разработанной модели на основе наборов больших данных.

Теоретическая и практическая ценность работы заключается в разработке теоретически корректной и экспериментально проверенной модели, реализующей базовые процедуры технологии машинного обучения. Программная реализация модели будет осуществлена на высокоуровневом языке программирования общего назначения и протестирована с использованием больших данных с возможностью повторного использования при решении различных задач, связанных с обработкой больших данных.

Сделан обзор литературы, изучение работ как отечественных, так и зарубежных ученных, посвященных тематике исследований в области анализа больших данных и методов машинного обучения. Определены основные методы исследования, такие как кластерный, компаративный, факторный анализ, методы машинного обучения, теория вероятностей и математической статистики. При разработке программного комплекса будет использован объектно-ориентированный подход.

По результатам исследования ожидается получение модели реализации алгоритм определения скрытых зависимостей между признаками данных, а также ее программная реализация.

Список литературы

1. BigDataStatistics 2020 //URL: https://techjury.net/stats-about/big-data-statistics/#gref

2. В мире подсчитано количество IoT-устройств //URL: http://www.dailycomm.ru/m/47373/

3. Искусственный интеллект (мировой рынок) //URL: http://www.tadviser.ru/index.php/Статья:Искусственный_интеллект_(мировой_рынок)

4. Переход на цифровое государство //URL: https://digitalkz.kz/perechod-na-cifrovoe-gosudarstvo/

5. Data mining //URL: https://ru.wikipedia.org/wiki/Data_mining

6. Fan J., Han F., and Liu H. Challenges of Big Data Analysis // Princeton University, Johns Hopkins University, August 7, 2013 //URL: https://arxiv.org/pdf/1308.1479.pdf

Научный руководитель, PhD Исмаилова А.А.