Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.855.5

Шлеменков
Алексей Андреевич

Решение прикладных задач методами машинного обучения

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Егорова Н.Г.
кандидат технических наук,
доцент

Минск 2019

КРАТКОЕ ВВЕДЕНИЕ

На данный момент у мирового сообщества почти не осталось сомнений по поводу того, что искусственный интеллект вместе с машинным обучением станут одной из центральных компонент в современных интеллектуальных системах. В связи с быстрым ростом вычислительной мощности компьютерных систем, наличием больших успехов в прикладных задачах: медицине, рекламе, банковской сфере, телекоммуникациях, стремительным развитием сообщества, подведением научных обоснований под методы машинного обучение, оно стал стало крайне популярным и полезным для множества компаний.

Системы машинного обучение на данный момент не только выполняют рутинные занятия (заменяя при этом человека), но и делают некоторые виды работ быстрее, качественнее, дешевле, чем люди. Примером такой задачи может послужить классификация изображений. Последние разработки в этой сфере позволили достичь рекордной точности, которая не только равна, но и превышает человеческую.

Несмотря на всеобщую популярность, машинное обучение не является панацеей от всех проблем и универсальным их решением. Внедрение машинного обучения в уже существующих системы требует высокого уровня организованности данных, и, естественно, достаточного их количества. Тем не менее, глубокое погружение в область задачи является практически обязательным условием построения качественной системы машинного обучения, направленной на решение бизнес-проблем. Поэтому в работе рассмотрено не только машинное обучение в целом, но и большое внимание уделено конкретной предметной области, в которой определена задача: обработка текста на естественном языке.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью данного исследования является поиск и применение современных подходов к анализу данных к конкретной исследуемой задаче, разработка моделей и алгоритмов машинного обучения и оценка их качества относительно выбранной задачи.

Для достижения поставленной цели необходимо решить следующие задачи:

1 Проанализировать современное состояние области машинного обучения и анализа данных;

2 Выбрать конкретную прикладную задачу и более детально проанализировать эту область знаний

3 Провести анализ и моделирование выбранной задачи. Провести эксперименты, сравнить результаты и сделать вывод о том, какими методами и моделями машинного обучения можно наилучшим образов (в заданных ограничениях) решить поставленную задачу.

Объектом исследования являются выбранная прикладная задача.

Предметом исследования — применение алгоритмов машинного обучения для решения поставленной задачи.

Основной *гипотезой*, положенной в основу диссертационной работы, является наличие вычислительных мощностей, которые имеют возможность установки программного обеспечения с открытым исходным кодом, а также наличие размеченной выборки необходимого размера.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии научно-техническими заданиями и планами работ. В качестве прикладной задачи была выбрана задача со следующей формулировкой “определение вопросов дубликатов”. В рамках исследовательской работы над выбранной задачей была изучена предметная область машинного обучения и получено современное представление о ней в области NLP. Также была разработана система определения вопросов-дубликатов на английском языке, а также проведено сравнение различных подходов к решению задачи: использование RNN и CNN для анализа текстов.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Н. Г. Егоровой, заключается в формулировке целей и задач исследования.

Опубликованность результатов диссертации

По теме диссертации опубликованы две печатные работы в сборниках трудов и материалов международных научных конференции.

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций автора и приложения. В первой главе сформулирована и поставлена задача. Во второй главе сделан обзор предметной области: мащинного обучения в общем, и обработки естественного языка и приложений в частности, выявлены основные существующие проблемы в рамках тематики исследования, показаны направления их решения. В третьей главе дан обзор используемых технологий и их возможностей. В последней (четвертой главе) описан анализ данных, этапы создания дополнительных характеристик и обучения моделей, а также сравнение их производительности.

Общий объем диссертации − 78 страниц. Работа содержит 6 формул и 44 рисунка. Библиографический список включает 36 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во введении рассмотрено современное положение машинного обучения в научном и практическом сообществе, приведены его достижения и конкретно обозначена будущая решаемая задача, обоснована актуальность работы.

В первой главе сформулирована и конкретно поставлена задача, а также обозначен возможный способ применения разрабатываемой системы.

Во второй главе сделан обзор предметной области: машинного обучения в общем, и обработки естественного языка и приложений в частности: от определения областей применения машинного обучения – в домен решаемой задачи – обработку естественного языка. В подразделах второй главы также рассмотрены базовые понятия и подходы к моделированию языка с помощью *n*-граммных моделей, пояснен способ работы word2vec, более подробно обсуждены подходы к анализу текста при помощи LSTM и CNN.

В третьей главе дан обзор используемых технологий, которые широко применяются в современной разработке систем, использующих машинное обучение: Python, NumPy, Tensorflow, Gensim, Matplotlib, Spacy и другие.

В последней (четвертой) главе дано подробное описание используемых данных, проведен их анализ и рассчитаны дополнительные характеристики данных, проведен ряд визуализаций на каждом из этапов исследования. В этой же главе показано создание двух типов моделей и их архитектура в контексте решаемой задачи, введены и пояснены метрики качества, которые использовались для оценки производительности итоговой модели. В работе было рассмотрено два типа моделей. Один из них за основу берет LSTM ячейку, которая предназначена для моделирования последовательностей. Очевидно, что к последовательностям относится и текст на естественном (в данном случае – английском) языке. Другой же тип моделей подразумевает использование CNN для анализа текстов.

В заключении приведены основные достигнутые результаты и возможные будущие шаги, направленные на развитие области исследования.

ЗАКЛЮЧЕНИЕ

В работе была затронута тема машинного обучения и его применимости к прикладным задачам. В качестве практической задачи была выбрана область по обработке естественного языка: определение вопросов-дубликатов по их тексту. Для того, чтобы как можно более качественно выполнить поставленную задачу были выполнены следующие шаги:

1 Изучена область машинного обучения, проведена оценка его текущих возможностей.

2 Проведено глубокое погружение в область, которая напрямую связана с решаемой прикладной задачей.

3 Проведен анализ данных задачи.

4 Проведено моделирование и оценено качество полученных моделей.

5 Произведено сравнение моделей.

6 Сделан вывод о том, какая модель наилучшим образом решает поставленную задачу.

Можно сформулировать следующий список достигнутых результатов:

– точность классификации на отложенной выборке составила 85%, метрика кроссэнтропии – 0,34;

– скорость определения дубликатов достигла порядка сотен/тысяч пар вопросов в секунду, что является качественно новым результатом, в особенности если проводить сравнение со скоростью, с которой человек принимает аналогичное решение.

Поставленная задача была достигнута с использованием методов, позволяющих анализировать тексты и выделять в них контекстные связи, ассоциации. Несмотря на хорошее качество решения задачи, существует еще более новые методы, которые, возможно, позволят улучшить качество модели. К ним относятся механизм внимания (attention) и использование претренированных языковых моделей.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Шлеменков, А. А. Использование алгоритмов машинного обучения в анализе текстов естественного языка на примере определения вопросов-дубликатов / А. А. Шлеменков, Я. О. Гусак // Дистанционное обучение – образовательная среда XXI века : материалы X международной научно-методической конференции (Минск, 7 - 8 декабря 2017 года). – Минск : БГУИР, 2017. – С. 205 - 206.

2-А. Шлеменков, А. А. Использование многоуровневой модели для эффективного управления дамбой и предсказания наводнений / А. А. Шлеменков, Я. О. Гусак // Информационные технологии и системы 2018 (ИТС 2018) = Information Technologies and Systems 2018 (ITS 2018) : материалы международной научной конференции, Минск, 25 октября 2018 г. / Белорусский государственный университет информатики и радиоэлектроники ; редкол. : Л. Ю. Шилин [и др.]. – Минск, 2018. – С. 260 - 261.