Министерство образования Республики Беларусь
Учреждение образования

Белорусский государственный университет
информатики и радиоэлектроники

УДК

Бартош
Владислав Иванович

МЕТОДЫ И АЛГОРИТМЫ РАСПОЗНАВАНИЯ ТИПОВ СТАТЕЙ

АВТОРЕФЕРАТ
диссертации на соискание степени магистра технических наук

по специальности 1-40 80 02 «Системный анализ, управление и обработка информации»

_____

Научный руководитель
Севернёв Александр Михайлович
кандидат технических наук, доцент

_____

Минск 2020

ВВЕДЕНИЕ

Прогресс в области микроэлектроники и информационных технологий обусловил широкое распространение обработки в реальном времени больших потоков данных. Например, многие простые операции повседневной жизни, такие как использование кредитной карты или телефона, требуют автоматизированного создания, анализа и обработки различных данных. Поскольку эти операции часто выполняются большим числом участников, необходимы распределенные и массовые потоки данных. Точно так же социальные сети содержат большое количество специфических сетевых и текстовых потоков данных. Поэтому актуальна проблема создания моделей и алгоритмов, позволяющих эффективно обрабатывать большие потоки данных, особенно в условиях ограниченных временных и других ресурсов.

В современном мире, большое количество задач решается программным способом. Сейчас не составляет никакого труда посчитать систему нелинейных уравнений или создать точный прогноз погоды. Задачи, которые считались трудновыполнимыми для человека ранее — теперь решает компьютер. Но существует ряд задач, которые не под силу компьютеру. Например, безуспешно пытаться требовать компьютер рассказать о разнице между восприятием искусства ребенком и взрослым. Для решения данной проблемы в 50-х годах 20-го века были изобретены искусственные нейронные сети.

Актуальность исследований в данной области подтверждается огромным количеством самых разнообразных практических применений исскуственных нейронных сетей.

Методы машинного обучения применяются в различных областях науки: от обучения распознавания рукописного текста до классификации различных видов рака. Для обучения нейронной сети необходимо большое количество информации, поскольку невозможно добиться высокой точности работы алгоритмов машинного обучения на достаточно малом количестве данных. Так, к примеру, для анализа изображений *Google* в качестве обучающего набора данных использовал информацию с видеохостинга *YouTube*. В случае решения задачи распознавания речи, в качестве обучающей выборки использовалась серия аудиоклипов с приложенными к ним описаниями. Первая представленная версия распознавания речи на основе нейронной сети содержала уровень ошибок, достигающий 25%, через три года результат был улучшен и составлял уже 8% ошибок.

В последние несколько лет наблюдается огромный интерес к искусственным нейронным сетям. Они применяются в самых различных областях: в медицине, физике, технике. Нейронные сети вошли в практику везде, где есть необходимость решить задачи прогнозирования, классификации или управления. Огромный успех применения исскуственных нейронных сетей можно охарактеризовать несколькими причинами: они позволяют воспроизводить чрезвычайно сложные зависимости и справляются с задачами высокой размерности.

Методы классификации текстов лежат на стыке двух областей – информационного поиска и машинного обучения. Их сходство состоит в способах представления самих документов и способах оценки качества алгоритмов. На сегодняшний день разработано большое количество методов и их различных вариаций для классификации текстов. Каждая группа методов имеет свои преимущества и недостатки, области применения, особенности и ограничения.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы магистерской работы. В последние годы искусственный интеллект (ИИ) широко применяется в различных областях человеческой жизни. Неотъемлемой частью ИИ является машинное обучение (МО). Оно настолько глубоко вошло в нашу жизнь, что мы даже не замечаем его важность в современном мире. Именно благодаря МО поисковая машина понимает, какие результаты показывать на ваш запрос. Мировые биржевые фонды используют алгоритмы МО для предсказания стоимости акций, тем самым экономя кучу денег их владельцам. Вышеизложенное подтверждает актуальность темы магистерской диссертации.

Цель и задачи исследования. Целью исследования является обзор основных методов классификации текстовой информации и разработка оптимального алгоритма распознавания типов статей.

Для достижения поставленной цели ставятся следующие задачи:

− определить основные понятия ИИ;

− дать понятие классификации текста;

− определить основные задачи, которые можно решить с помощью классификации текста;

− описать задачу классификации текста;

− рассмотреть основные методы классификации текстовой информации;

− сформировать модель распознавания типов статей

− изучить алгоритм распознавания типов статей;

− выбор оптимального языка программирования для поставленной задачи;

− реализация алгоритма распознавания типов статей;

− продемонстрировать работу алгоритма.

Данные задачи последовательно решаются в главах диссертации.

Объект исследования. Текстовая информация в виде новостных статей.

Предмет исследования. Методы и алгоритмы распознавания типов статей.

Личный вклад соискателя. Основные результаты, изложенные в диссертации, получены автором самостоятельно. Научному руководителю в совместных работах принадлежат предметные постановки задач, выбор направлений исследования и анализ результатов.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Магистерская диссертация представлена в виде пояснительной записки на 46 страницах, состоящей из введения, четырёх разделов и заключения.

В первом разделе была рассмотрена модель классификации текста. Процесс классификации делится на два этапа: обучения и эксплуатации. На этапе обучения корпус документов преобразуется в векторы признаков. Затем признаки документов вместе с их метками (категориями или классами, распознаванию которых мы хотим обучить модель) передаются в алгоритм классификации, который определяет свое внутреннее состояние и выявленные шаблоны. После обучения можно векторизовать новый документ в то же пространство признаков и передать результат алгоритму прогнозирования, который вернет метку категории документа.

Во втором разделе были рассмотрены следующие методы классификации:

- вероятностные (метод Байеса);
- метрические (метод *k* ближайших соседей);
- логические (метод деревьев решений);
- линейные (метод опорных векторов);
- методы на основе искусственных нейронных сетей.

Были выделены преимущества и недостатки алгоритмов.

В третьем разделе осуществлялся выбор алгоритма распознавания типов статей. Была рассмотрена иерархическая сеть внимания как метод распознавания типов статей. Описывалась архитектура нейронной сети, при помощи которой будет реализован алгоритм распознавания типов статей.

В четвёртом разделе была рассмотрена реализация алгоритма распознавания типов статей на основании иерархической сети внимания. В ходе тестирования было выявлено, что более точных результатов можно добится на больших количествах данных.

ЗАКЛЮЧЕНИЕ

В ходе исследования, проведенного в данной работе, были получены следующие результаты.

Определены основные свойства и понятия нейронных сетей. Проанализировано применение нейронных сетей и машинного обучения в частности в повседневной жизни человека.

Затем ознакомились с понятием классификации текстовой информации, определили его математическую модель. Выявили основные недостатки линейного персептрона.

Были проанализированы существующие алгоритмы распознавания текстовой информации. Были выделены преимущества и недостатки алгоритмов.

Проанализировав существующие алгоритмы распознавания текстовой информации, была выбрана иерархическая сеть внимания как метод распознавания типов статей. Была рассмотрена архитектура нейронной сети, при помощи которой будет реализован алгоритм распознавания типов статей.

В алгоритмах глубокого обучения точность классификации существенно зависит от наличия обучающей выборки подходящего размера. Подготовка такой выборки – очень трудоемкий процесс. Следует отметить, что обучение нейронной сети проводилось на коллекциях англоязычных текстов.

Для демонстрации работы алгоритма мы воспользовались набором данных из новостного портала *huffpost.com*, по которым мы определяли тип статьи. Для реализации алгоритма был использован язык программирования *Phyton* и его основные модули для работы с данными.

Таким образом, был разработан алгоритм распознавания типов статей с применением иерархической сети внимания, которая позволяет решать задачи классификации различной степени сложности. Степень ошибки алгоритма были сведены к минимуму, но полностью не удалось избавиться от них.

Результаты работы были доложены на 56-ой научной конференции аспирантов, магистрантов и студентов Белорусского государственного университета информатики и радиоэлектроники в виде доклада на тему «Методы распознавания типов статей».

СПИСОК ПУБЛИКАЦИЙ АВТОРА

[1–А.] Бартош, В.И. Методы распознавания типов статей / В.И. Бартош // Информационные технологии и управление: материалы 56-й научной конференции аспирантов, магистрантов и студентов. (Минск, 21 – 24 апреля 2020 г.). – Минск: БГУИР, 2020. – С.71–72.