МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

ИССЛЕДОВАНИЯ ДАННЫХ ЭКОНОМИЧЕСКОЙ ПРИРОДЫ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Мукомел Александра Юрьевича

Научный руководитель:
Старший преподаватель _____ Е.Е. Лапшева
подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева
подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Машинное обучение, основанное на анализе данных, зародилось около 70 лет назад. Инструменты в этой предметной области уже давно используются для решения широкого ряда научных задач, таких как:

- исследование далекого космоса;
- открытие новых частиц;
- анализ последовательностей ДНК;
- разработка индивидуальных методов лечения болезней и мн. др.

И несмотря на то, что за прошедшие годы общие принципы не изменились, бурный рост вычислительных мощностей компьютеров, наряду с глобальной цифровизацией, позволил многократно усложнить закономерности и прогнозы, которые можно получить, используя методы машинного обучения. Также увеличился круг решаемых ими проблем и задач, и машинное обучение стало использоваться не только в крупномасштабных или меняющих мир задачах. Круг коммерческих задач, которые можно решить с помощью машинного обучения довольно обширен:

- оценка платёжеспособности по кредиту;
- прогнозирование покупательского спроса;
- вычисление логистических цепочек и мн. др.

А в последние годы применение методов машинного обучения в повседневной жизни стало обыденностью. Многие современные веб-сайты и устройства используют алгоритмы машинного обучения, начиная с автоматических рекомендаций по просмотру фильмов, заказа еды или покупки продуктов, и заканчивая распознаванием друзей на фотографиях. Практически любой сложный сайт, такой как Яндекс, Google или Кинопоиск содержит несколько моделей машинного обучения.

Часто, говоря о ML, подразумевают нейронные сети, но это не совсем так, ведь нейронные сети лишь одна из множества методик анализа данных. Тем не менее это объяснимо тем, что нейросети в последние годы находятся на

пике своей популярности. По данным Google Trends пик поисковых запросов с упоминанием «Deep Learning» приходится на 2018 год (см. рисунок 1).

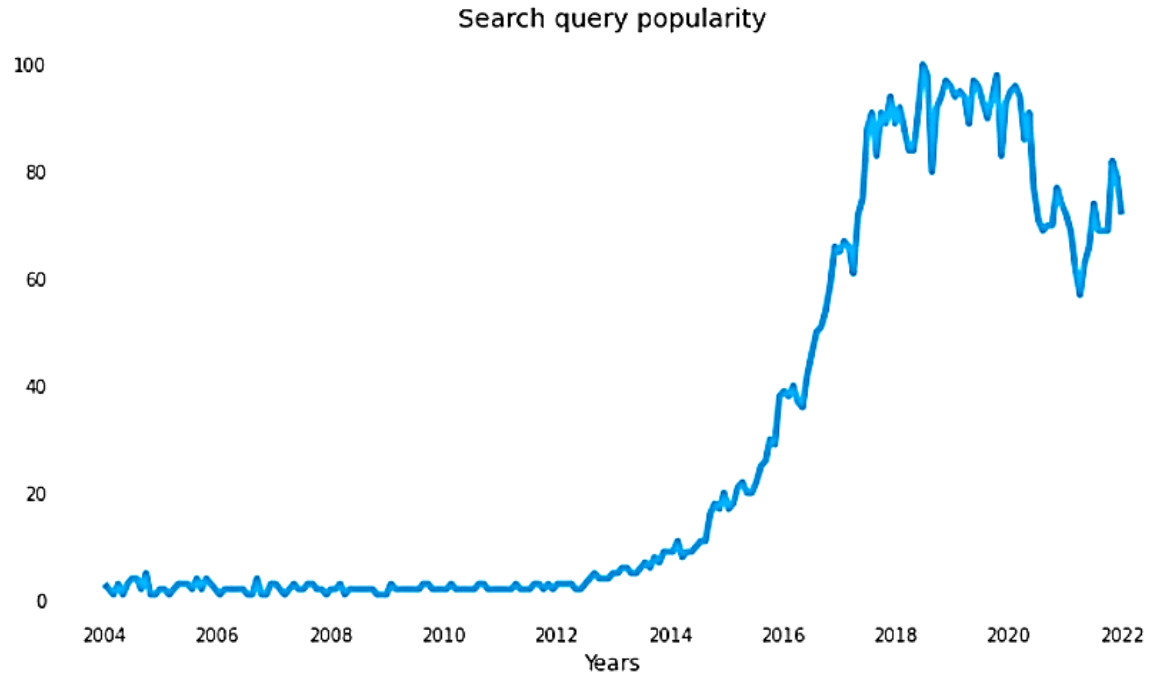

Рисунок 1 – график популярности поисковых запросов с упоминанием «Deep Learning» среди всех стран.

Числа на графике обозначают уровень интереса к теме по отношению к наиболее высокому показателю в таблице для определенного периода времени. Т.е. 100 баллов – наивысший уровень популярности запроса, 50 – уровень популярности запроса, вдвое меньший по сравнению с первым случаем.

То, что произошло с нейросетями – пример того, как использование давно описанной теории было недоступно ввиду технических ограничений, а именно:

- высокой сложности сбора большого набора данных необходимого для эффективного обучения;
- очень высокой длительностью обучения из-за низких мощностей компьютеров.

Сегодня же существует множество открытых площадок для сбора данных, изображений, текста, аудио и видеосигналов для решения сложных задач, а мощные графические процессоры, и усовершенствование

математической составляющей процесса обучения нейросетей позволяют расширить область их применения. В результате с помощью нейросетей стало возможным обрабатывать данные огромных объемов, такие как:

- пакеты, состоящие из миллионов изображений;
- данные временных рядов за годы наблюдений;
- измерения, параллельно собранные с тысяч датчиков и мн. др.

Наиболее интересны для рассмотрения временные ряды. Потому что, в отличие от задач поперечного сечения для их анализа и прогнозирования имеет значение порядок измерений, т.к. любой временной ряд по определению представляет собой набор упорядоченных во времени значений некоторой переменной.

Цель бакалаврской работы – прогнозирование поведения временного ряда экономической природы с использованием методов машинного обучения.

Поставленная цель определила **следующие задачи**:

1. Ознакомиться с основными методами анализа и прогнозирования временных рядов, и выбрать наиболее подходящий.
2. Изучить подходящие области ML.
3. Найти источник для сбора данных, необходимых для анализа.
4. Изучить программные средства необходимые для реализации.
5. С помощью выбранного метода реализовать программу для построения прогностической модели.
6. Протестировать и настроить модель.
7. Оценить точность выбранной модели.
8. Проанализировать результаты работы.

Методологические основы машинного обучения представлены в работах А. Мюллера, У.С. Мак-Каллока, С. Хохрайтера. М. Тэдди, Т. Фергюсона, С. Рудера и др.

Теоретическая и/или практическая значимость бакалаврской работы. Выбор и применение эффективных методов машинного обучения для

прогнозирования поведения временных рядов экономической природы позволит игрокам на финансовых рынках придерживаться правильной стратегии.

Структура и объём работы. Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и 7 приложений. Общий объем работы – 78 страниц, из них 63 страницы – основное содержание, включая 40 рисунков и 2 таблицы, список использованных источников информации – 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Использование машинного обучения при прогнозировании поведения временных рядов» посвящен теоретическим основам методов машинного обучения, которые можно использовать для прогнозирования временных рядов экономической природы.

В первом разделе даются определения следующим понятиям и терминам:

Машинное обучение – это методики анализа данных, которые позволяют аналитической системе обучаться и учиться выявлять закономерности в ходе решения множества сходных задач с минимальным участием человека.

Машинное обучение с учителем – объединение видов машинного обучения, заключающихся в обучении алгоритма на предоставленных ему парах объект-ответ, в результате чего алгоритм получает способность выдать ответ для ранее не встречавшегося объекта, без какой-либо помощи человека

Искусственная нейронная сеть – математическая модель и её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей – сетей нервных клеток живого организма.

Перцептрон – простейший вид нейронных сетей, предложенный Ф. Розенблаттом в 1958 году, в основе которого лежит модель восприятия информации мозгом, реализованная в виде электронной машины «Марк-1» в 1960 году.

Рекуррентные нейронные сети – модели искусственных нейронных сетей, в которых возможно использовать петли обратной связи.

Длинная цепь элементов краткосрочной памяти – разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям.

Оптимизаторы – важный компонент архитектуры нейронных сетей, определяющий оптимальный набор параметров модели, таких как вес и

смещение, чтобы при решении конкретной задачи модель выдавала наилучшие результаты.

Стохастический градиентный спуск – один из самых популярных методов для оптимизации целевой функции с подходящими свойствами гладкости.

Алгоритм оптимизации Адама – расширение алгоритма стохастического градиентного спуска, который в последнее время широко используется в приложениях глубокого обучения, особенно для таких задач, как компьютерное зрение и обработка естественного языка.

Проблема переобучения – модель хорошо обрабатывает только примеры из обучающей выборки, адаптируясь к обучающим примерам, вместо того чтобы учиться классифицировать не участвовавшие в обучении примеры, т.е. теряет способность к обобщению.

Исключение – метод регуляризации искусственных нейронных сетей, предназначенный для уменьшения переобучения сети за счет предотвращения адаптации совокупности нейронов на тренировочных данных во время обучения.

Машинное обучение без учителя – объединение видов машинного обучения, заключающихся в предоставлении алгоритму только объектов, из которых необходимо извлечь знания, т.е. отсутствует указывающий ответ алгоритму учитель.

Нормализация – преобразование данных к неким безразмерным единицам, для приведения данных в разных единицах измерения и диапазонах значений к единому виду, который позволит сравнивать их между собой или использовать для расчёта схожести объектов.

На основе рассмотренных методик анализа временных рядов и их особенностей были выбраны наиболее интересные и эффективные, что послужило фундаментом для практической части работы.

Второй раздел «Методики и средства для анализа временных рядов экономической природы» посвящен реализации нейронной сети для прогнозирования временных рядов.

Во втором разделе была реализована модель рекуррентной нейронной сети LSTM с использованием стандартных средств языка Python и следующих библиотек: Keras, Scikit-learn, NumPy, Matplotlib, Pandas. После настройки модели был проведен ряд экспериментов при ее различных конфигурациях, на основании которых была проведена оценка результатов работы модели.

Исходя из данных, полученных при обучении и прогнозировании стоимости акций компаний Apple, Amazon и Tesla был сделан вывод, что максимальная точность модели достигается при ее обучении на данных компании, для которой осуществляется прогнозирование. При этом ее обобщающая способность применительно к другим компаниям хуже, но приемлемая. В случае обучения модели на наборах данных различных компаний ее обещающие способности повышаются, но вместе с этим может падать точность прогноза для конкретных компаний.

ЗАКЛЮЧЕНИЕ

Семейство алгоритмов, известное как нейронные сети, недавно пережило свое возрождение под названием «глубокое обучение». Несмотря на то, что глубокое обучение сулит большие перспективы в различных сферах применения машинного обучения, алгоритмы глубоко обучения, как правило, жестко привязаны к конкретным случаям использования.

Область применения машинного обучения безгранична и, учитывая все многообразие данных, имеющихся на сегодняшний день, ограничивается лишь воображением. Данная работа также подтверждает, что в настоящее время для реализации модели ГНС не требуется дорогостоящего оборудования и для получения первого опыта и результатов в этой области может быть достаточно идеи и энтузиазма.

Подводя итоги проведенной работе, можно сделать вывод: поставленная цель была выполнена, в ходе работы были рассмотрены различные методики машинного обучения для анализа и прогнозирования данных, изучены методики глубинного обучения и программные средства необходимые для реализации модели нейронной сети и обработки данных (библиотеки Keras, Scikit-learn, NumPy, Matplotlib, Pandas).

В практической части работы была реализована и настроена прогностическая модель рекуррентной нейронной сети LSTM, была проведена оценка результатов ее работы. LSTM – одна из наиболее часто используемых нейронных сетей для анализа временных рядов. Способность LSTM запоминать предыдущую информацию делает его идеальным для таких задач.

Основные источники информации:

1. Мюллер, А. Введение в машинное обучение с помощью Python – руководство для специалистов по работе с данными / А. Мюллер, С. Гвидо – М.: Б. и., 2016-2017 – 393 с.
2. Мак-Каллок, У. Логическое исчисление идей, относящихся к нервной активности / У.С. Мак-Каллок, В. Питтс; ред. К.Э. Шеннон, Д. Маккарти – М.: Иностранная литература, 1956 – 363-384 с.
3. Hochreiter, S. Long short-term memory / Neural Computation Vol. 9, no. 8 [journal] / S. Hochreiter; J. Schmidhuber – Cambridge: MIT Press, 1997 – 33 с.
4. Taddy, M. Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions / M. Taddy – New York: McGraw-Hill, 2019 – 352 с.
5. Ferguson, T. An inconsistent maximum likelihood estimate / Journal of the American Statistical Association [journal] / T. S. Ferguson – 1982. — Т. 77, вып. 380. – JSTOR 2287314, с. 831–834
6. Robbins, H. A convergence theorem for non-negative almost supermartingales and some applications / Optimizing Methods in Statistics / H. Robbins, D.O. Siegmund; ed. J.S. Rustagi – Academic Press, 1971 – 421-436 с.
7. Ruder, S. An overview of gradient descent optimization algorithms / Insight Centre for Data Analytics [article] / S. Ruder, Dublin: NUI Galway, 2016 – 2-8 с.
8. Srivastava, N. Dropout: A Simple Way to Prevent Neural Networks from Overfitting / Journal of Machine Learning Research [journal] / N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever [et al.]; ed. Yoshua Bengio – Toronto: University of Toronto, 2014
9. Freedman, D. Statistics: Fourth International Student Edition / D. Freedman, R. Pisani, R. Purves – New-York: W.W. Norton & Company, 2007 – 701 с.