УДК 004 **Информационные технологии**

Капустин Алексей Ильич
студент МГТУ им. Н.Э. Баумана, г. Москва
e-mail: akapust1n@mail.ru

МЕТОД ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ТЕННИСНЫХ МАТЧЕЙ НА ОСНОВЕ АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация: В работе проведен краткий обзор современных подходов к прогнозированию результатов теннисных матчей. Предложен новый метод прогнозирования на основе анализа естественного языка. Описан набор данных для тестирования алгоритма, а также меры по его очистке и нормализации. Проведено тестирование метода с различным набором параметров: различные наборы словарей, разное количество эпох обучения нейронной сети, различные функции потерь. Предложены направления дальнейшего улучшения алгоритма.

Ключевые слова: прогнозирование, нейронные сети, естественный язык.

Annotation: The paper provides a brief overview of modern approaches to predicting the results of tennis matches. A new forecasting method based on natural language analysis is proposed. The data set for testing the algorithm is described, as well as measures for its cleaning and normalization. The method was tested with a different set of parameters: different sets of dictionaries, different numbers of neural network learning eras, different loss functions. The directions of further improvement of the algorithm are proposed.

Keywords: forecasting, neural networks, natural language.

Введение. Прогнозирование результатов спортивных событий является

сложной задачей, особенно в случае спорта, демонстрирующего весьма стохастический характер. Например, в футболе отслеживаются многочисленные особенности, которые объединяются с экспертными знаниями, что дает возможность создания и использования многочисленных различных алгоритмов прогнозирования. Большинство современных алгоритмов прогнозирования основаны на анализе статистики: истории предыдущий матчей, мест в рейтингах и других различных статистических факторов. Нет реальной возможности предсказать точно результат события, но используя методы прогнозирования, можно снизить неопределенность. Современная наука пытается использовать любую информацию о событии для составления прогнозов. В некоторых видах спорта составители соревнований перед/после матча устраивают пресс-конференции, на которых спортсмены общаются с прессой. Эту информацию потенциально можно использовать для прогнозирования. В данной работе будет произведен анализ возможности применения нейронных сетей, обученных на словах спортсменов с пресс-конференций.

Современные методы прогнозирования. Существуют два основных подхода к прогнозированию результатов спортивных событий – прогнозирования результатов соревнования в целом и прогнозирование результатов конкретного матча. В свою очередь методы в подходах делятся на следующие классы: основные на статистике, экспертные и экспортно-статистические методы. В статистический методах изучается набор данных с некоторым количеством показателей, из которых могут выбираться наиболее значимые. Экспертные методы основаны на знаниях и опыте высококвалифицированных специалистов в рассматриваемой предметной области.

Набор данных для анализа. Для анализа был выбран наиболее тривиальный вид спорта для прогнозирования - одиночный теннис. Два спортсмена, два возможных исхода матча. Были собраны стенограммы теннисных пресс-

конференций от ASAP Sports [1], чей архив стенограмм тенниса начинается 1992 году и до сих пор регулярно обновляется. Для нашего исследования мы берем интервью перед игрой для теннисных одиночных матчей, сыгранных между январем 2000 года и октябрем 2015 года. Стенограммы сопоставляются с исторической спортивной статистикой с сайта Tennis-Data, который охватывает большинство профессиональных мужских матчей, сыгранных в 2000-2015 годах и женских в 2007-2015. В результате сопоставления этих наборов данных, был получен датасет из следующего вида: результат матча - интервью данное после предыдущего матча. Все интервью предоставлены на английском языке.

```
1|I played well When I figured it out I gave myself.
1|No the conditions are very changing during
1|Yeah a lot's different
1|I lost a little bit in a part
1|No I served well
```

Рисунок 1. Пример набора данных

Из набора данных были удалены вопросы журналистов - оставлены только ответы спортсменов на них. Удалены все небуквенные символы (кроме пробелов), буквы переведены в единый регистр.

Метод прогнозирования.

Создание модели прогнозирования происходит в 2 этапа

- Формирование "словаря" - ограниченного набора слов с указанием частоты их нахождения в наборе данных для обучения
- Обучение нейронной сети. При обучении в процессе векторизации используется словарь, созданный на первом этапе

```
model.add(Dense(512, input_shape=(max_words,) ,
 activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(256, activation='sigmoid'))
model.add(Dropout(0.5))
model.add(Dense(2, activation='softmax ' ))
```

Рисунок 2. Псевдокод нейронной сети

Нейронная сеть включает в себя 5 слоев - 3 плотных(так называемых dense layer) и два слоя потерь (позволяют бороться с переобучением сети).

Набор данных был разбит две части - тестовая и обучающая часть(в соотношению 9 к 1), при этом тестовые данные не использовались при обучении. Было проведено измерение численных характеристик точности(Precision) и полноты(Recall) в зависимости от количества эпох при размере словаря 4000. Из измерения можно сделать вывод, что оптимальным количеством эпох для данной архитектуры сети является 5.

Так же было протестировано изменение размера словаря при фиксированном количестве эпох. Оптимальным является размер словаря - 4000, при увеличении размера словаря точность не растет, а при 6000 начинается падать.

Таблица 1. Результаты обучения нейронной сети в зависимости от количества эпох

Количеств	Точ	По
3	0.68	1
4	0.72	1
5	0.75	1
6	0.7	1
7	0.65	0.9
8	0.58	1
9	0.63	0.9
10	0.62	1

Таблица 2. Результаты обучения нейронной сети в зависимости от размера словаря

Количеств	Точ	По
400	0.5	1
1000	0.62	0.8
2000	0.7	0.9
3000	0.71	1
4000	0.75	1
5000	0.75	1
6000	0.73	1

Так же было проведено тестирование изменения функции потерь при обучении нейронной сети при размере словаря 4000 и 5-ти эпохах обучения. Из вышеприведенных измерений можно.

Таблица 3. Результаты обучения нейронной сети в зависимости от функции потерь

Количество эпох	Точ	По
categorical_crossentropy	0.72	1
sparse_categorical_crossen	0.62	1
binary_crossentropy	0.75	1
kullback_leibler_diverge	0.62	0.9
poisson	0.65	1
cosine_proximity	0.5	1

Выводы. На относительно небольшом объеме данных(2100 записей) для исследования было показано, что слова спортсменов на пресс- конференциях можно использовать для прогнозирования результатов событий. Полученная модель умеет точность пред- сказаний примерно 0.75, что является высоким результатом, по сравнению с другими аналогичными работами [2, с. 620]. Недостатком данного метода можно считать, что при работе с другими наборами данных для обучения и прогнозирования, точность прогнозирования может измениться.

Улучшение результатов прогнозов. В схожих научных статьях и дипломах иностранных вузов [3, с. 264; 4, с. 65] исследования схожих тем проводились на выборках 7000-33000 записей. Поэтому для улучшения результатов прогнозирования можно сделать следующие вещи:

- Увеличить размер набора данных
- Сравнить различные архитектуры нейронных сетей
- Улучшить очистку текста: убрать падежи, числа, местоимения.

Библиографический список:

1. База данных спортивных интервью / ASAP Sports Interviews[Электронный ресурс] URL: http://www.asapsports.com/recent.php (Дата обращения 03.03.2020).

2. Morton A., McHale I. A Bradley-Terry type model for forecasting tennis match results. / International Journal of Forecasting. – 2011. – №27. – С.620.

3. Newton P.K., Keller J.B. Probability of Winning at Tennis I. Theory and Data / Studies in Applied Mathematics. – 2005. - №114(3) – С.241-269.

4. Peters J. Predicting the Outcomes of Professional Tennis Matches. / University of Edinburgh. – 2017. – С. 64-66.