Моделирование и анализ данных
2023. Том 13. № 2. С. 64–76
DOI: https://doi.org/10.17759/mda.2023130203
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)
© 2023 ФГБОУ ВО МГППУ

Modelling and Data Analysis
2023. Vol. 13, no. 2, pp. 64–76
DOI: https://doi.org/10.17759/mda.2023130203
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)
© 2023 Moscow State University of Psychology & Education

УДК 004.85

Проблемы классификации текстов естественного языка методами классического машинного обучения

Сологуб Г.Б.*
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-5657-4826
e-mail: glebsologub@ya.ru

Пухов В.А.**
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-8078-6386
e-mail: csguard26@gmail.com

В статье описаны проблемы методов классического машинного обучения в задаче классификации естественного языка. Одной из таких задач является классификация структурных элементов в школьных сочинениях. На её примере рассматриваются недостатки классического машинного обучения по сравнению с другими, более сложными алгоритмами.

Ключевые слова: классификация текста, анализ естественного языка, автоматизация проверки сочинений.

Для цитаты:
Сологуб Г.Б., Пухов В.А. Проблемы классификации текстов естественного языка методами классического машинного обучения // Моделирование и анализ данных. 2023. Том 13. № 2. С. 64–76. DOI: https://doi.org/10.17759/mda.2023130203

***Сологуб Глеб Борисович,** кандидат физико-математических наук, доцент кафедры математической кибернетики института «Компьютерные науки и прикладная математика» Московский авиационный институт (национальный исследовательский университет) (МАИ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-5657-4826, e-mail: glebsologub@ya.ru

****Пухов Вячеслав Александрович,** студент магистратуры института «Компьютерные науки и прикладная математика», Московский авиационный институт (национальный исследовательский университет) (МАИ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-8078-6386, e-mail: csguard26@gmail.com

1. ВВЕДЕНИЕ

Задача классификации текста является одной из базовых задач анализа естественного языка.

Проблема выбор метода классификации текста возникла при решении задачи сегментации текста школьных сочинений в рамках соревнования по машинному обучению. Для проверки сочинения требуется выделить в тексте структурные элементы, такие как введение, позиция автора, аргумент и т.д.

При решении данной задачи была предложена модель машинного обучения, которая сначала разделяет структурные элементы, затем их классифицирует.

В этой статье анализируются и сравниваются методы классификации текстов на описанных выше данных.

Для обработки естественного языка был использован метод TF-IDF, впервые изложенный в [1]. Этот метод считает частоту вхождения слов в подстроке и взвешивает их в отношении к частоте встречаемости этого слова в документе. Таким образом, более релевантные слова получают больший вес в векторном представлении текста. Однако, такой способ не учитывает порядок слов в подстроке.

Описание алгоритмов машинного обучения и метрики их качества взяты из [2], что позволило выбрать множество методов для анализа, критерий эффективности работы методов для задачи классификации структурных элементов школьных сочинений, а также такой способ сравнения алгоритмов как обучающая и валидационная кривые.

В работах [3] и [4] описывается архитектура сети LSTM, выбранной в качестве исследуемого метода глубокого обучения.

2. ПОСТАНОВКА ЗАДАЧИ

Имеется датасет (Таблица 1) из 1487369 строк и 3 столбцов.

Таблица 1

id	discourse_text	discourse_type
423A1CA112E2	Modern humans today are always on their phone...	Lead
423A1CA112E2	They are some really bad consequences when stu...	Position
423A1CA112E2	Some certain areas in the United States ban ph...	Evidence
…	…	…
4C471936CD75	it is better to seek multiple opinions instead.	Position
4C471936CD75	The impact of asking people to help you make a...	Evidence
4C471936CD75	there are many other reasons one might want to…	Concluding Statement

Столбец id – идентификатор сочинения, discourse_text – текст структурного элемента, discourse_type – тип структурного элемента.

Возможные типы структурных элементов:

- введение (Lead);
- позиция автора сочинения (Position);
- аргумент (Claim);
- контраргумент (Counterclaim);
- опровержение контраргумента (Rebuttal);
- пример, подтверждающий аргумент (Evidence);
- вывод (Concluding Statement).

Этот набор данных описывает текстовые документы, содержащие сочинения. Каждая строка содержит id документа, подстроку соответствующего документа, выделенную как структурный элемент, и тип этого элемента.

Задача классификации формулируется так: пусть X – множество описаний объектов, Y – множество меток классов. Существует неизвестная целевая зависимость – отображение $y^*: X \to Y$, значения которого известны только на объектах данной обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Требуется построить алгоритм $a: X \to Y$, способный классифицировать произвольный объект $x \in X$.

В контексте данной работы структурные элементы из данного набора данных будут рассматриваться без привязки к текстовым документам: входные признаки X^m – тексты структурных элементов, и целевые метки классов Y^m – типы структурных элементов.

Задача этой работы заключается в сравнении нескольких алгоритмов a. Целью работы ставится выявление наиболее эффективного алгоритма для классификации структурных элементов школьных сочинений с точки зрения заданной метрики.

3. ПРЕДОБРАБОТКА ТЕКСТА

Для работы с естественным языком были произведены преобразования исходного текста структурных элементов в векторные представления:

- количественная векторизация,
- TF-IDF.

Количественная векторизация представляет собой преобразование входного текста к матрице, где номер каждого столбца – индекс слова из словаря входного текста, номер каждой строки – порядковый номер предложения, а значение элемента матрицы – количество вхождений соответствующего слова в соответствующем предложении.

TF-IDF трансформация состоит из:

1. TF (term frequency – частота слова) – отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова t в пределах отдельного документа

$$tf(t, d) = \frac{n_t}{\sum_k n_k},$$

где n_t есть число вхождений слова t в документ, а в знаменателе – общее число слов в данном документе.

2. IDF (inverse document frequency – обратная частота документа) – инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов

$$idf\left(t,D\right)=\frac{\log\left|D\right|}{\left|\left\{d_i\in D\middle|t\in d_i\right\}\right|},$$

где $\left|D\right|$ – число документов в коллекции, $\left|\left\{d_i\in D\middle|t\in d_i\right\}\right|$ – число документов из коллекции D, в которых встречается t (когда $n_t\neq 0$).

3. $TF\text{ - }IDF\left(t,d,D\right)=tf\left(t,d\right)\times idf\left(t,D\right)$.

Формулы $TF\text{ - }IDF$ приведены в главе 6 [1].

После того, как тексты преобразованы в векторную форму, можно применять алгоритмы классификации.

4. СРАВНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ

При оценке качества алгоритмов машинного обучения использовалась метрика Macro F1 Score [2]:

$$Macro\,F1Score=\frac{\sum_{i=1}^{N}F1Score_i}{N},$$

где N – число классов структурных элементов (7); i – номер класса; $F1Score_i=\dfrac{TP_i}{TP_i+\dfrac{1}{2}\left(FP_i+FN_i\right)}$, TP_i – количество верно классифицированных объектов класса i; FP_i – количество объектов неверно отнесенных к классу i; FN_i – количество объектов класса i, неверно отнесенных к другому классу.

Таким образом, $Macro\,F1Score$ – суть среднее арифметическое $F1Score$ по каждому классу. В свою очередь $F1Score$ – среднее гармоническое точности и полноты классификации.

Подобный выбор метрики обусловлен тем, что в задаче сегментации текста сочинений нет предпочтения ложноположительным или ложноотрицательным ошибками, а количество структурных элементов в обучающем множеств – несбалансированное.

Для анализа эффективности алгоритмов использовались обучающая и валидационная кривые. Первая позволяет понять, как влияет мощность обучающей выборки на оценочную метрику. Вторая, методом скользящего контроля – какое значение гиперпараметра оптимально для решения задачи. Также, для сохранения времени исходный датасет был сокращен до 100000 структурных элементов.

Классические модели машинного обучения подразделены на метрические и неметрические. Одни используют векторное расстояние между объектами, другие – нет.

5. КЛАССИЧЕСКИЕ НЕМЕТРИЧЕСКИЕ АЛГОРИТМЫ

Наиболее совершенным классическим алгоритмом машинного обучения является градиентный бустинг. При исследовании задачи классификации рассматривались несколько его вариаций на деревьях решений: стохастический градиентный бустинг (sklearn GradientBoostingClassifier), экстремальный градиентный бустинг (XGBoost) и CatBoost.

Сравнение алгоритмов представлено в таблице 2 [2].

Таблица 2

	GBC	**XGBoost**	**CatBoost**
Построение деревьев	По уровням	По уровням	По уровням однородно
Поиск расщеплений	Полный перебор или гистрограммный подход	Полный перебор или гистрограммный подход	Предварительный биннинг (дискретизация вещественных признаков)
Важность признаков	Impurity	Gain / Frequency или Weight / Coverage	Изменение прогнозируемых значений / функции ошибки
Ранняя остановка	+	-	+

6. СТОХАСТИЧЕСКИЙ ГРАДИЕНТНЫЙ БУСТИНГ

На рисунках 1 и 2 представлены соответственно обучающая и валидационная кривые для стохастического градиентного бустинга.

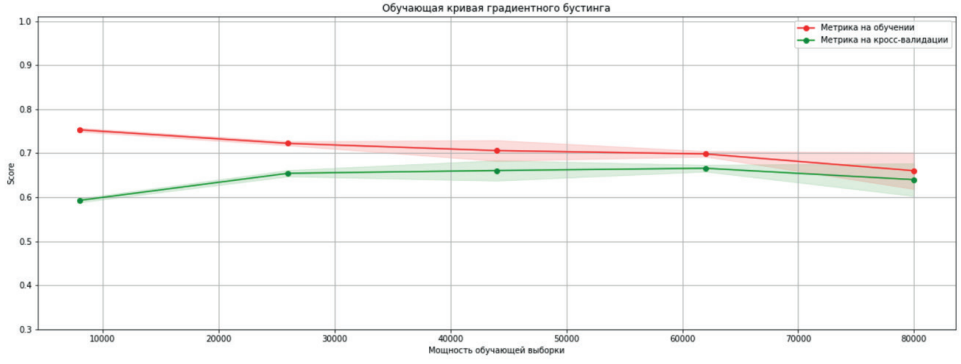

Рис. 1. Обучающая кривая метода GBC

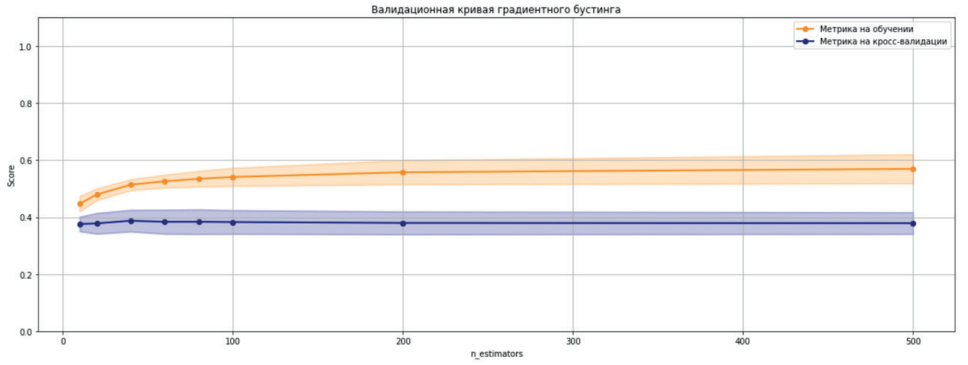

Рис. 2. Валидационная кривая метода GBC

XGBoost

На рисунках 3 и 4 представлены соответственно обучающая и валидационная кривые для экстремального градиентного бустинга.

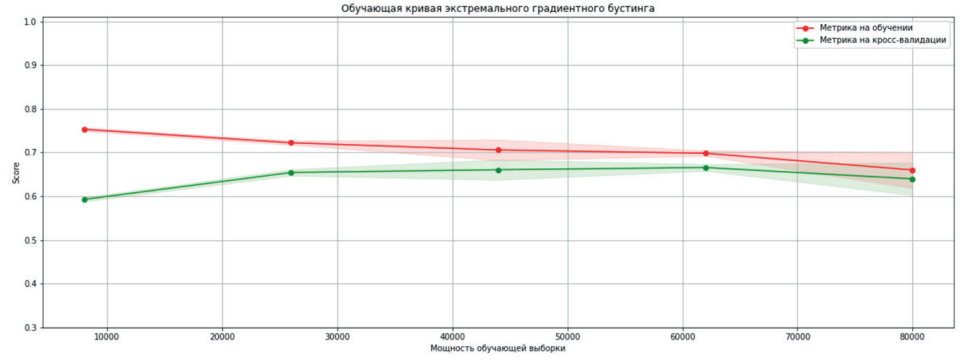

Рис. 3. Обучающая кривая метода XGBoost

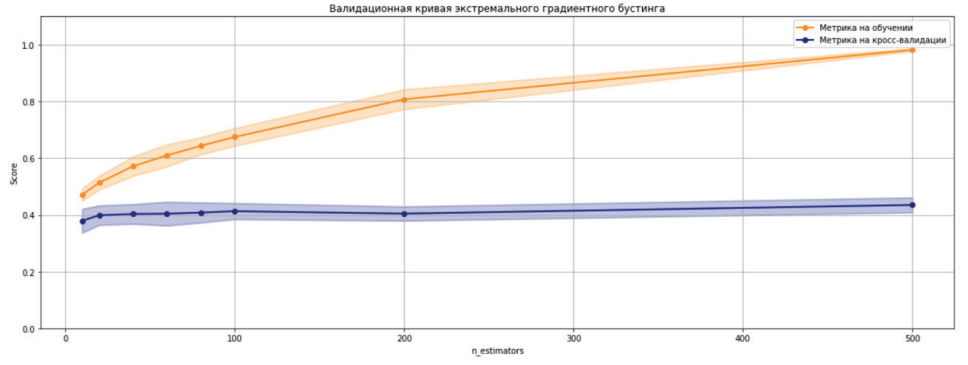

Рис. 4. Валидационная кривая метода XGboost

CatBoost

На рисунках 5 и 6 представлены соответственно обучающая и валидационная кривые для метода CatBoost.

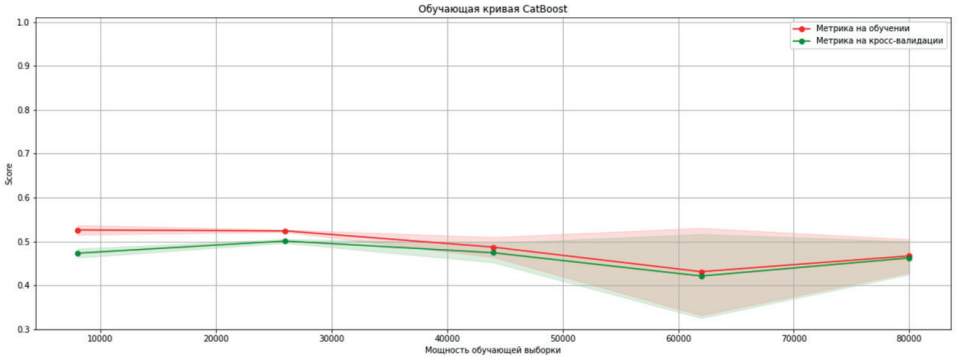

Рис. 5. Обучающая кривая метода CatBoost

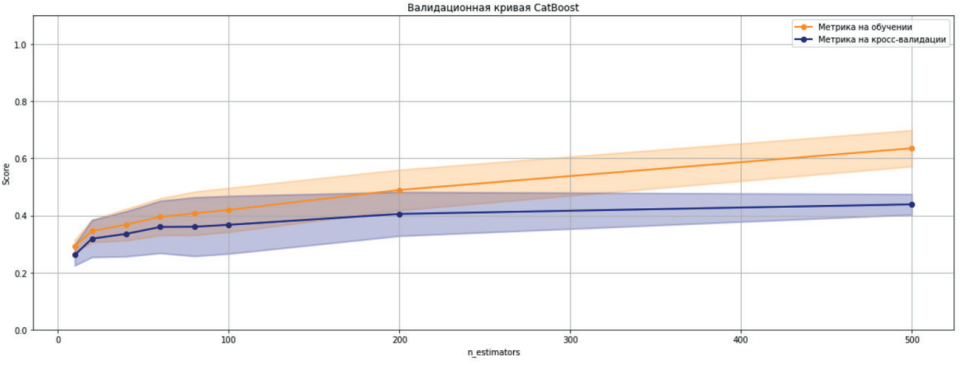

Рис. 6. Валидационная кривая метода CatBosst

Анализ обучающих кривых показывает, что первые две модели перестают улучшаться при увеличении мощности обучающего множества более 70000 образцов, CatBoost имеет аналогичную картину, за исключением необычно большого разброса значения метрики на предпоследнем значении мощности тренировочной выборки.

Валидационные кривые свидетельствуют о том, что увеличение сложности моделей не влияют на целевую метрику.

Максимальное значение метрики на отложенной выборке при решении задачи классификации градиентным бустингом составило $Macro\,F1Score = 0.668$.

7. КЛАССИЧЕСКИЕ МЕТРИЧЕСКИЕ АЛГОРИТМЫ

Метрические модели, в основном, работают без учителя, поэтому анализировался только метод ***N* ближайших соседей**, обучающийся на данных.

На рисунках 7 и 8 представлены соответственно обучающая и валидационная кривые для метода *N* ближайших соседей.

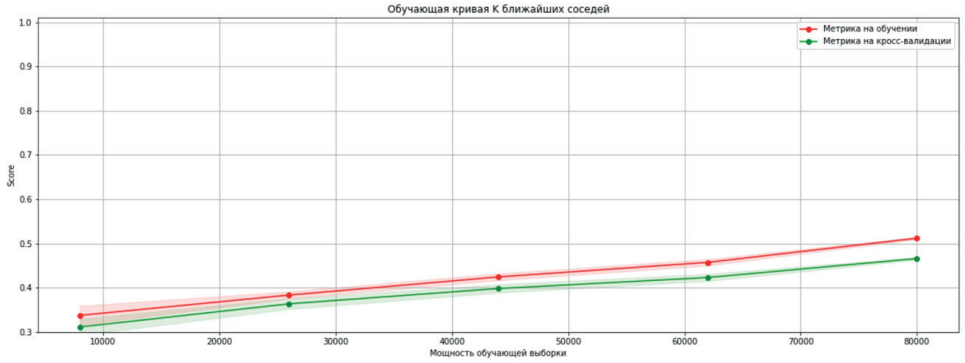

Рис. 7. Обучающая кривая метода N ближайших соседей

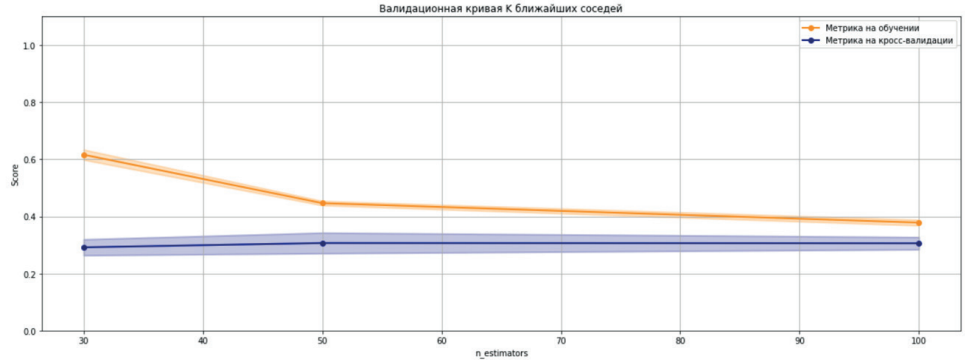

Рис. 8. Валидационная кривая метода N ближайших соседей

В случае *N* ближайших соседей, обучающая кривая имеет тренд на увеличение, но само обучение замедляется с ростом обучающей выборки, так как алгоритм считает расстояние между всеми её объектами. Поэтому, учитывая выход на плато кривой валидации, этот метод не оказался более эффективным, чем бустинг.

8. ГЛУБОКОЕ ОБУЧЕНИЕ

Далее, был исследован подход глубокого обучения. Наиболее эффективной из рассмотренных оказалась нейросеть долгой краткосрочной памяти (LSTM) [3].

LSTM – рекуррентная нейронная сеть, способная удалять информацию из состояния ячейки.

Пусть x_t – t-ое входное значение фрагмента A нейронной сети, h_t – возвращаемое значение (рисунок 9).

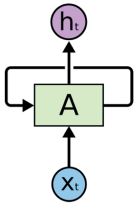

Рис. 9

LSTM слой состоит из нескольких подслоев [4].

1. Фильтр забывания (рисунок 10). Сигмоидальный слой, возвращающий значение от 0 до 1, отвечающий за сохранение информации из ячейки C_{t-1}.

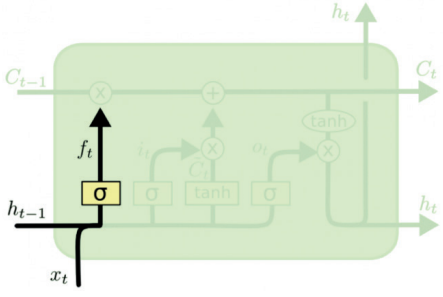

Рис. 10. Схема фильтра забывания

$$f_t = \sigma\left(W_f\left[h_t - 1, x_t\right] + b_f\right),$$

где W_f – веса слоя, b_f – вектор смещения.

2. Сигмоидальный и *tanh* слой изображен на рисунке 11. Решает, какая новая информация будет храниться в состоянии ячейки.

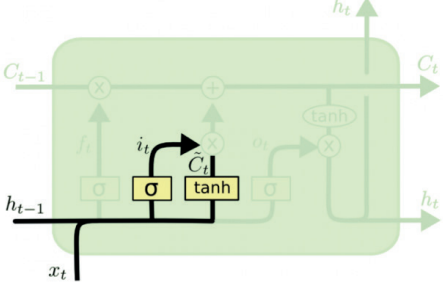

Рис. 11. Схема обновления информации

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right)$$

$$\widetilde{C_t} = \tanh\left(W_C\left[h_{t-1}, x_t\right] + b_C\right),$$

где W_i, W_C и b_i, b_c – веса и смещения соответствующих индексу слоёв, C_t – вектор значений новых кандидатов на добавление в состояние ячейки.

3. Этап замены состояния C_{t-1} на $C_t = f_t C_{t-1} + i_t \widetilde{C_t}$ изображен на рисунке 12.

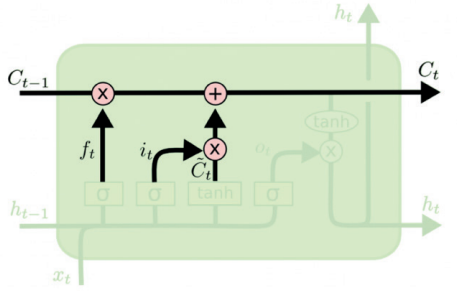

Рис. 12. Схема замены состояния ячейки

4. Получение выходного значения. Схема показана на рисунке 13.

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t \cdot \tanh\left(C_t\right),$$

W_o – веса выходного слоя, b_o – выходной вектор смещения.

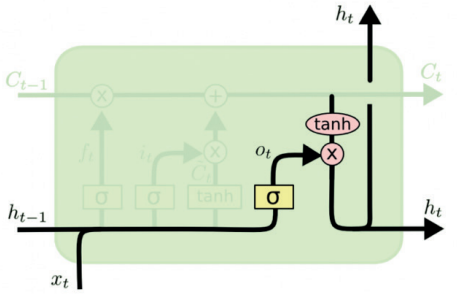

Рис. 13. Схема получения выходного значения ячейки

Структура полученной сети:

1) эмбеддинговый слой,
2) LSTM-слой из 128 нейронов,
3) слой прореживания,
4) LSTM-слой из 64 нейронов,

5) слой прореживания,
6) softmax слой.

Валиационная кривая и зависимость функции ошибки от итерации обучения построенной нейронной сети изображены на рисунке 14.

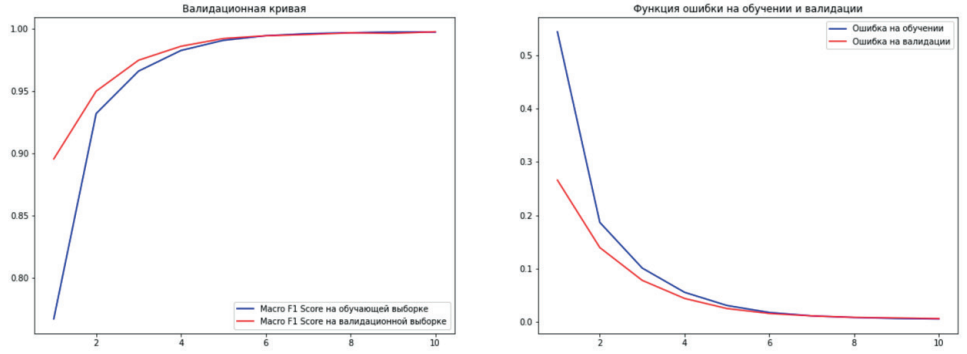

Рис. 14. Валидационная кривая и функция ошибки нейронной сети

Судя по полученным графикам, сеть полностью выучивает обучающую выборку. Тем не менее, значения метрики на отложенной выборке не убывает с каждой итерацией обучения.

Качество такой сети зависит от того, насколько тестовые примеры будут отличаться от данных в датасете. Возможно, следует понизить количество итераций обучения, если для тестовых данных исходная выборка будет не репрезентативна. Исходя из полученного значения метрики $Macro\,F1Score = 0.973$ нейронные сети эффективнее решают задачу классификации текстов в сравнении с моделями классического машинного обучения.

9. ЗАКЛЮЧЕНИЕ

При анализе методов классического машинного обучения (Gradient Boosting, CatBoost, XGBoost, *N* ближайших соседей) и метода глубокого обучения (LSTM нейронной сети) для решения задачи классификации структурных элементов текстов школьных сочинений было выявлено, что, с точки зрения выбранной метрики, лучше справился метод глубокого обучения.

Литература

1. *Manning, C.D.; Raghavan, P.; Schutze, H.* Scoring, term weighting, and the vector space model // Cambridge University Press. 2009 P. 109–133 DOI:10.1017/CBO9780511809071.007
2. *Дьяконов А.Г.* Лекции [Электронный ресурс] URL: https://dyakonov.org/tag/лекции/
3. *Alex Sherstinsky* Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network [Электронный ресурс] // Physica D: Nonlinear Phenomena 2020

P. 1–40 DOI:10.1016/j.physd.2019.132306 URL: https://sciencedirect.com/science/article/abs/pii/S0167278919305974

4. *Christopher Olah* Understanding LSTM Networks [Электронный ресурс] // 2015 URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Problems of Natural Language Classification Using Methods of Classical Machine Learning

Gleb B. Sologub*
Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia
ORCID: https://orcid.org/0000-0002-5657-4826
e-mail: glebsologub@ya.ru

Vyacheslav A. Pukhov**
Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia
ORCID: https://orcid.org/0000-0002-8078-6386
e-mail: csguard26@gmail.com

This article describes the problems of classical machine learning methods in natural language classification. One of these tasks is the classification of structural elements in school essays. On its example, the shortcomings of classical machine learning are considered in comparison with other, more complex algorithms.

Keywords: text classification, natural language analysis, automation of essay checking.

For citation:
Sologub G.B., Pukhov V.A. Problems of Natural Language Classification Using Methods of Classical Machine Learning. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2023. Vol. 13, no. 2, pp. 64–76. DOI: 10.17759/mda.2023130203 (In Russ., abstr. in Engl.).

References

1. Manning, C.D., Raghavan, P., Schutze, H. Scoring, term weighting, and the vector space model // Cambridge University Press. 2009 P. 109–133 DOI:10.1017/CBO9780511809071.007
2. Dyakonov A.G. Lectures *https://dyakonov.org/tag/лекции/ (*In.Russ,*)*
3. Alex Sherstinsky Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network // 2020 P. 1–40 DOI:10.1016/j.physd.2019.132306 Available at: https://sciencedirect.com/science/article/abs/pii/S0167278919305974 *(*In.Russ)
4. Christopher Olah Understanding LSTM Networks // 2015 Available at: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

***Gleb B. Sologub,** Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Mathematical Cybernetics, Institute of Computer Science and Applied Mathematics, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0002-5657-4826, e-mail: glebsologub@ya.ru

****Vyacheslav A. Pukhov,** Student of the Institute of Computer Science and Applied Mathematics, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0002-8078-6386, e-mail: csguard26@gmail.com

Получена 21.03.2023 Received 21.03.2023
Принята в печать 21.04.2023 Accepted 21.04.2023