Методы машинного обучения для классификации текстовой информации

Н.А. Кривошеев[1], В.Г. Спицын[1]
nikola0212@mail.ru | spvg@tpu.ru
[1]Национальный исследовательский Томский политехнический университет, Томск, Россия

Рассматривается метод классификации текстовой информации на основе аппарата сверточных нейронных сетей. Приведен алгоритм предобработки текста. Предобработка текста состоит из: лемматизации слов, удаления стоп-слов, обработки символов текста и т.д. Производится пословное преобразование текста в плотные вектора. Тестирование проводится на базе текстовых данных «The 20 Newsgroups». Данная выборка содержит коллекцию примерно из 20 000 новостных документов на английском языке, которая разделена (приблизительно) равномерно между 20 различными категориями. Точность лучшей из применявшихся в данной работе сверточной нейронной сети на тестовой выборке составила ~ 74%. Приведена топология лучшей нейронной сети. Точность голосования нейронных сетей алгоритмом Бэггинга составила ~ 81.5%. На основе проведенного обзора аналогичных решений приведено сравнение со следующими алгоритмами классификации текста: методом опорных векторов (SVM, 82.84%), наивным байесовским классификатором (81%), алгоритмом k ближайших соседей (75.93%), мешком слов.

Ключевые слова: *нейронный сети, Бэггинг, классификация текста, база данных «The 20 Newsgroups».*

Machine Learning Methods for Classification Textual Information

N.A. Krivosheev[1], V.G. Spitsyn[1]
nikola0212@mail.ru | spvg@tpu.ru
[1]National Research Tomsk Polytechnic University, Tomsk, Russia

A method for classifying textual information based on the apparatus of convolutional neural networks is considered. The text preprocessing algorithm is presented. Text preprocessing consists of: lemmatizing words, removing stop words, processing text characters, etc. The word-by-word conversion of the text into dense vectors is performed. Testing is carried out on the basis of the text data of "The 20 Newsgroups". This sample contains a collection of approximately 20,000 news stories in English, which is divided (approximately) evenly between 20 different categories. The accuracy of the best convolutional neural network used in this work on the test set was ~ 74%. The topology of the best neural network is given. The accuracy of voting of neural networks by the Bagging algorithm was ~ 81.5%. Based on a review of similar solutions, a comparison is made with the following text classification algorithms: the support vector method (SVM, 82.84%), the naive Bayes classifier (81%), the k nearest neighbors algorithm (75.93%), and the word bag.

Keywords: *neural networks, Bagging, text classification, database "The 20 Newsgroups".*

1. Введение

На данный момент одной из наиболее популярных задач является понимание текста. К данной задаче относятся: классификация, перевод, ответы на вопросы и др. Задача классификации является одной из традиционных в машинном обучении, в связи с чем были созданы базы текстовых данных для обучения нейронных сетей.

Существует множество решений задачи классификации текстов [1, 3, 5, 12], которые используют различные методы преобразования текста в вектора, такие как: посимвольное преобразование, N-граммы символов, преобразование слов, Word2vec, мешок слов и другие. Существует множество алгоритмов для классификации текстов: нейронные сети, метод опорных векторов (SVM), k ближайших соседей, наивный байесовский классификатор и др.

В данной работе рассматривается задача классификации текстов с помощью многослойного персептрона и сверточной нейронной сети. Рассмотрена предобработка текстовых данных в виде пословного преобразования текста в вектора. Рассмотрено голосование нейронных сетей методом Бэггинга [11]. Приведены результаты обучения и тестирования нейронных сетей на тестовой выборке «The 20 Newsgroups» [14, 15], проведено сравнение с аналогами. Все программы реализованы на языке Python, с использованием библиотеки Keras.

Далее будут рассмотрены возможные решения задачи классификации текстов с описанием результатов тестирования на базе текстовых данных «The 20 Newsgroups» [14, 15]. На основе проведенного обзора аналогичных решений приведено сравнение со следующими алгоритмами классификации текста: методом опорных векторов (SVM, 82.84%), наивным байесовским классификатором (81%), алгоритмом k ближайших соседей (75.96%), мешком слов.

2. Обучающая выборка

Для обучения, тестирования и сравнения с аналогами была выбрана обучающая выборка «The 20 Newsgroups» [14, 15]. Данная выборка содержит коллекцию примерно из 20 000 новостных документов на английском языке, которая разделена (приблизительно) равномерно между 20 различными категориями. К данным классам относятся такие темы как: религия, наука, политика, спорт и др. Коллекция «The 20 newsgroups» стала популярным набором данных для экспериментов с техниками машинного обучения для текстовых приложений, таких как классификация текста.

3. Обоснование выбора сверточной нейронной сети

Сверточная нейронная сеть (СНС) в последние годы становится все более популярной и используется в решении различных задач. СНС хорошо показала себя в решении задач, связанных с обработкой естественного языка. Данный алгоритм является очень гибким и может использоваться для классификации с использованием различных методов предобработки текста. СНС классифицирует текстовые данные значительно лучше многослойного персептрона [2].

Основной особенностью СНС является использование фильтров чувствительных к определенной последовательности слов.

Предшественниками свёрточных нейронных сетей были модели когнитрона и неокогнитрона. Свёрточные нейронные сети в современном виде были представлены в работах Ле Куна [7-9] и A. Krizhevsky, I. Sutskever, G.E. Hinton [6]. Основными слоями, используемыми в свёрточных сетях, являются: свёрточный слой, слой субдискретизации и полносвязный слой.

Сочетание функции классификации с функцией выделения признаков с помощью ядер свёртки, получаемых в процессе обучения, позволяет выделять оптимальный набор признаков. Получить данный набор признаков, подбирая метод извлечения признаков вручную, является практически невозможной задачей.

Исходя из вышесказанного, было выдвинуто предположение, что аппарат сверточных нейронных сетей может показать высокую эффективность в решении задачи классификации текстовых данных.

Полученные результаты необходимо сравнить с другими подходами, не использующими нейронные сети, по точности решения поставленной задачи.

4. Предобработка данных

Первичная предобработка заключается в переводе текста в нижний регистр, в удалении малоинформативных и редких символов. К редким символам относятся символы, используемые во всей выборке не более нескольких десятков раз. Проводится замена на пробелы следующих символов: табуляция, перевод строки, длинная последовательность одинаковых символов (например, множество из трех и более звездочек). Проводится удаление стоп-слов. Производится лемматизация [13] всех слов в тексте (см. рис. 1).

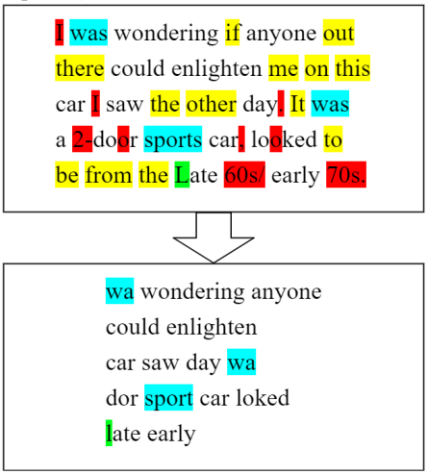

Рис. 1. Пример обработки текста.

В данной работе производится пословное преобразование текста в вектор. Проводится подсчет количества упоминаний каждого слова. Очень редкие и короткие слова удаляются. Составляется словарь слов, в котором каждому слову назначается индивидуальное число для последующей подачи на нейронную сеть.

Следует отметить, что точность сверточных нейронных сетей существенно зависит от составленного словаря, так если не провести удаление относительно редких слов или неправильно подобрать порог частоты упоминания, то точность сети значительно снижается и может снизиться почти в 2 раза. В данной работе порог частоты упоминания слова в тексте подбирается экспериментальным путем. Если слово упоминается меньшее число раз, чем в каждом 25*N тексте (где N - количество классов), то оно удаляется. Данная формула выведена экспериментально для данной выборки и является ее приближением. При изменении обучающей выборки может потребовать корректировки.

Следующим этапом предобработки является преобразование текста в вектор, в данной работе используется автоматическое преобразование в плотные вектора фиксированного размера. Данное преобразование производится за счет слоя Embedding библиотеки Keras.

После предобработки все тексты стандартизируется (обрезаются или заполняются) до заданной длины. В данной работе все тексты стандартизованы до длины текста в 300 слов. Если длина текста меньше 300 слов, то недостающая часть вектора заполняется нулями.

5. Результаты обучения и тестирования нейронных сетей

В данной работе протестированы сверточные нейронные сети различных топологий. Был применен алгоритм Бэггинга.

Все топологии нейронных сетей обучались с помощью метода NADAM [10] с использованием категориальной функции потерь (categorical_crossentropy). Во всех скрытых слоях нейронной сети используется функция активации RELU [4]. Выходной слой использует функцию активации softmax [16].

Алгоритм Бэггинга использует голосование 7-ми сверточных нейронных сетей. Голосование проходит путем поэлементного перемножения выходных векторов нейронных сетей.

Топология лучшей из применявшихся в данной работе сверточной нейронной сети приведена на рис. 2.

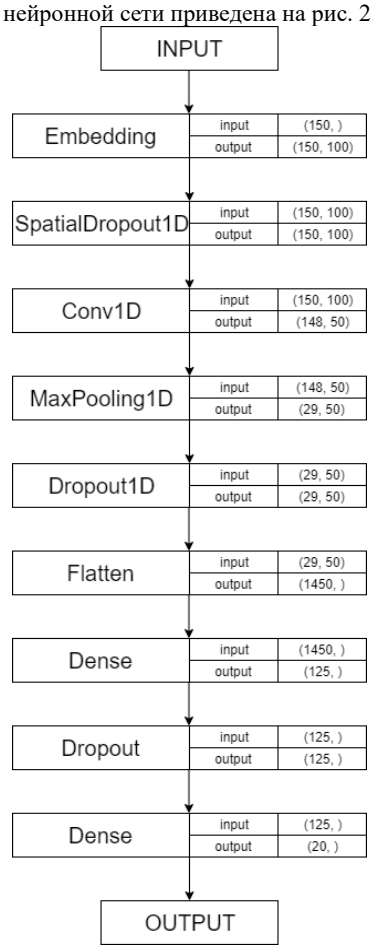

Рис. 2. Топология сверточной нейронной сети, лучшей из применявшихся в данной работе.

Топологии сверточных нейронных сетей различны (количество слоев, число нейронов в слое, размер окна свертки и т.д.).

В данной работе было проведено тестирование нейронных сетей на тестовой выборке «The 20 Newsgroups» [14, 15] с использованием пословного преобразования текста в плотные вектора.

Точность распознавания лучшей используемой сверточной нейронной сети на тестовой выборке составляет ~74%. Средняя точность используемых нейронных сетей составляет ~71.9%. При использовании голосования сверточных нейронных сетей методом Бэггинга, точность классификации возрастает до ~81.5%.

На основании полученных данных можно сделать вывод, что алгоритм Бэггига значительно повысил точность классификации.

6. Сравнение с аналогами

Существует множество аналогов для решения задачи классификации текстов, которые были протестированы на базе текстовых данных «The 20 Newsgroups» [1, 3, 5]. В представленных аналогах используются такие методы как: многослойный персептрон, сверточные нейронные сети, наивный байесовский классификатор, метод опорных векторов (SVM).

Методы предобработки текста в аналогах могут различаться между собой и не совпадать с методом предобработки текста в данной статье, который был описан выше.

В статье [3] автор использует мешок слов и многослойный персептрон в качестве классификатора. В указанной статье автор использует три класса из обучающей выборки: comp.graphics, sci.space, rec.sport.baseball. Точность нейронной сети, используемой в статье [3] составила 75%.

В данной работе было проведено тестирование голосования нейронных сетей (алгоритма, описанного выше) на данных указанных в статье [3], точность на тестовой выборке составила ~95.5%. Точность голосования сверточных нейронных сетей значительно превышает точность алгоритма, указанного в статье [3]. Сравнение точности алгоритмов представлено на рис. 3:

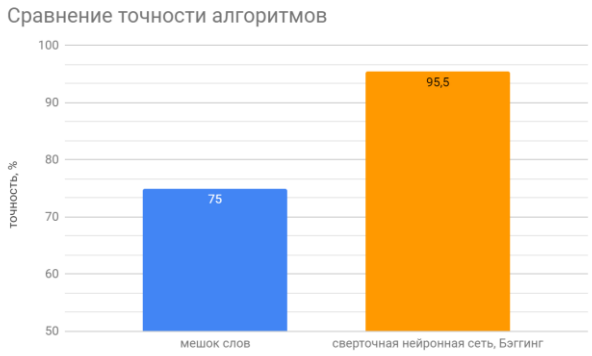

Рис. 3. Сравнение точности алгоритма с аналогом из статьи [3].

В статье [5] автор провел тестирование следующих алгоритмов: метод опорных векторов (SVM), k ближайших соседей, наивный байесовский классификатор и др. В статье автор использует все 20 классов из обучающей выборки. Точность классификатора, основанного на методе опорных векторов, составляет 82.84%, данный алгоритм является лучшим из представленных в статье [5] при обучении и тестировании на выборке «The 20 Newsgroups» [14, 15]. Точность метода опорных векторов превышает точность классификатора, основанного на голосовании сверточных нейронных сетей. Точность наивного байесовского классификатора, представленного в статье [5] составляет 81%, что значительно превосходит алгоритм k ближайших соседей, который был протестирован в статье [5]. Точность алгоритма k ближайших соседей составляет 75.93%.

В данной работе было проведено тестирование голосования нейронных сетей методом Бэггинга на данных указанных в статье [5], точность на тестовой выборке составила ~81.5%. Сравнение точности алгоритмов представлено на рис. 4:

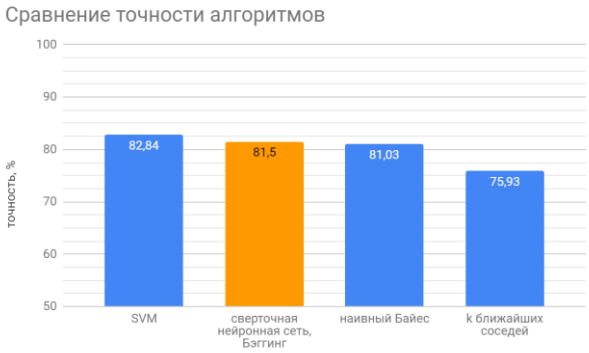

Рис. 4. Сравнение точности алгоритма с аналогами из статьи [5].

В статье [1] автор провел тестирование следующих алгоритмов: наивный байесовский классификатор и метод опорных векторов (SVM). В указанной статье автор использует четыре класса из обучающей выборки: alt.atheism, comp.graphics, sci.med, soc.religion.christian. Точность наивного байесовского классификатора, используемого в статье [1] составила 83.4%. Точность метода опорных векторов (SVM), используемого в статье [1] составила 91.2%.

В данной работе было проведено тестирование голосования нейронных сетей методом Бэггинга на данных указанных в статье [1], точность на тестовой выборке составила ~92%, что незначительно превосходит метод опорных векторов (SVM, 91.2%). Сравнение точности алгоритмов представлено на рис. 5:

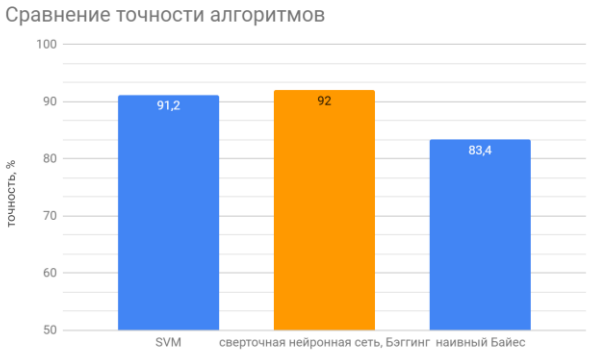

Рис. 5. Сравнение точности алгоритма с аналогами из статьи [1].

На основе проведенного сопоставления с аналогами можно сделать вывод, что голосование нейронных сетей методом Бэггинга может конкурировать с такими методами как наивный байесовский классификатор и метод опорных векторов (SVM).

7. Заключение

В данной работе реализовано и протестировано голосование сверточных нейронных сетей алгоритмом Бэггига. На основе полученных результатов можно сделать вывод, что голосование сверточных нейронных сетей с использованием Бэггинга показало существенный рост

точности классификации по сравнению с ранее полученными результатами [2] и может конкурировать с другими представленными алгоритмами, предназначенными для классификации текста. В дальнейшем планируется поиск новых и улучшение используемых способов решения данной задачи.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-08-00977 А.

8. Литература

[1] Веревкина О. Работа с текстовыми данными в scikit-learn [Электронный ресурс]. — Режим доступа: URL: https://habr.com/ru/post/264339/ (20.05.2019).

[2] Кривошеев Н.А., Спицын В.Г. Алгоритмы понимания текста методами глубокого обучения нейронных сетей // Сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых «Молодежь и современные информационные технологии» – Томск, 2018 г., с. 82-83.

[3] Мескита Д. Общий взгляд на машинное обучение: классификация текста с помощью нейронных сетей и TensorFlow [Электронный ресурс]. — Режим доступа: URL: https://tproger.ru/translations/text-classification-tensorflow-neural-networks/ (21.11.2018).

[4] Петренко С. Это нужно знать: Ключевые рекомендации по глубокому обучению (Часть 2) [Электронный ресурс]. — Режим доступа: URL: http://datareview.info/article/eto-nuzhno-znat-klyuchevyie-rekomendatsii-po-glubokomu-obucheniyu-chast-2/ (20.05.2019).

[5] Cardoso A. Datasets for single-label text categorization [Электронный ресурс]. — Режим доступа: URL: http://ana.cachopo.org/datasets-for-single-label-text-categorization (03.06.2019).

[6] Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. 2012, pp. 1097–1105.

[7] LeCun Y. Backpropagation applied to handwritten zip code recognition // Neural computation. 1989, Vol. 1(4), pp. 541–551.

[8] LeCun Y., Bottou L., Bengio Y., Haffner P. Gradientbased learning applied to document recognition // Proceedings of the IEEE. 1998, Vol. 86(11), pp. 2278-2324.

[9] LeCun Y. Efficient backprop // Neural Networks: Tricks of the Trade: Lecture Notes in Computer Science / G. Montavon, G. B. Orr, K.-R. Muller (Eds.) – Springer, 2012, pp. 9-48.

[10] Ruder S. An overview of gradient descent optimization algorithms [Электронный ресурс]. — Режим доступа: URL: http://ruder.io/optimizing-gradient-descent/index.html#nadam (22.11.2018).

[11] Бэггинг [Электронный ресурс]. — Режим доступа: URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%91%D1%8D%D0%B3%D0%B3%D0%B8%D0%BD%D0%B3 (25.08.2019).

[12] Классификация текста с помощью нейронной сети на Java [Электронный ресурс]. — Режим доступа: URL: https://habr.com/post/332078/ (21.11.2018).

[13] Лемматизация [Электронный ресурс]. — Режим доступа: URL: https://dic.academic.ru/dic.nsf/ruwiki/1313114/%D0%9B%D0%B5%D0%BC%D0%BC%D0%B0%D1%82%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F (26.08.2019).

[14] 20 Newsgroups [Электронный ресурс]. — Режим доступа: URL: http://qwone.com/~jason/20Newsgroups/ (10.09.2019).

[15] sklearn.datasets.fetch_20newsgroups [Электронный ресурс]. — Режим доступа: URL: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html (22.11.2018).

[16] Softmax [Электронный ресурс]. — Режим доступа: URL: https://medium.com/@congyuzhou/softmax-3408fb42d55a (20.05.2019).

Об авторах

Кривошеев Николай Анатольевич, магистрант отделения информационных технологий Национального исследовательского Томского политехнического университета. E-mail: nikola0212@mail.ru.

Спицын Владимир Григорьевич, д.т.н., профессор отделения информационных технологий Национального исследовательского Томского политехнического университета. E-mail: spvg@tpu.ru.