Инженерный вестник Дона, №1 (2025)
ivdon.ru/ru/magazine/archive/n1y2025/9741

Влияние уменьшения размерности словоформенных эмбеддингов на качество классификации текста

А.А. Сайгин, С.А. Федосин

Мордовский государственный университет им. Н.П. Огарёва, Саранск, Россия

Аннотация: В статье представлены существующие методы уменьшения размерности данных для обучения машинных моделей естественного языка. Вводятся понятия векторизации текста и слоформенного эмбеддинга. Формируется задача классификации текста. Формируются этапы обучения классификатора. Проектируется классифицирующая нейронная сеть. Проводится серия экспериментов на определение влияния уменьшения размерности словоформенных эмбеддингов на качество классификации текста. Сравниваются результаты оценки работы обученных классификаторов.

Ключевые слова: обработка естественного языка, векторизация, словоформенный эмбеддинг, классификация текста, уменьшение размерности данных, классификатор.

Классификация – понятие в науке, обозначающее разновидность деления объёма понятия по некоторому признаку или критерию. В процессе классификации объём понятия делится на виды, виды – на подвиды и т.д. [1]. Классификация широко применяется в практической деятельности, особенно в науке. Задача классификации – задача присвоения объектам значения класса. Имеется конечное множество объектов, для которых известно, к каким классам они относятся – выборка. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества. Классифицировать объект – значит, указать наименование класса, к которому относится данный объект [2].

Одной из более узких задач классификации является классификация текстов и документов. Данная задача относится к области обработки естественного языка. Классификация документов находит применение для решения многих задач: фильтрации спама, контекстной рекламы, в поисковых системах.

Как и в любой другой задаче обработки естественного языка, основная сложность при классификации текста состоит в том, что компьютер оперирует числовыми значениями, а не текстовыми. Для решения этой проблемы было разработано множество алгоритмов векторизации, отображающие текст в словоформенные эмбеддинги. Принцип работы каждого алгоритма различается, из-за чего каждый обладает своими достоинствами и недостатками [3].

В настоящее время для классификации применяются алгоритмы машинного обучения. На вход классификатора подается набор признаков в виде числового вектора, на выходе получается значение класса. С увеличением количества признаков усложняется архитектура обучаемого алгоритма, что приводит к увеличению объёма итоговой модели и времени ее обучения, ухудшению результативности. В задачах машинного обучения в целом и классификации в частности для решения данной проблемы прибегают к уменьшению размерности данных. Методы уменьшения размерности предоставляют меньшее количество измерений при сохранении наиболее важной информации. При этом могут потеряться некоторые детали, но это компенсируется более простым представлением данных, которое легче обрабатывать и сравнивать [4].

При классификации текстов в роли признаков выступают словорменные эмбеддинги. Каждый метод векторизации текста дает на выходе векторы большой размерности, что осложняет процесс классификации. Поэтому можно попробовать уменьшить их размерность. Для большинства задач обработки естественного языка такие векторы станут бесполезными, но задачи классификации это касается в меньшей степени.

Методы уменьшения размерности данных делятся на два типа: линейные и нелинейные. Линейные методы фиксируют исходные закономерности в данных, за счет чего представляют данные в пространстве

меньшего размера. Нелинейные методы фиксируют более сложные нелинейные зависимости в признаках, и так же используют их для уменьшения размерности [4].

Пример метода уменьшения размерности – анализ главных компонент (Principal Component Analysis, PCA). Данный метод уменьшает размерность набора данных, максимизируя при этом дисперсию каждого основного компонента. Дисперсия характеризует степень колебания значений столбца. Отсюда, объекты с большей дисперсией содержат большую информацию, объекты с нулевой дисперсией не несут информации. PCA ищет сжатое представление данных в меньшем измерении, максимизирующего общую дисперсию исходных данных [2].

Еще один метод уменьшения размерности, часто применяемый на практике – анализ независимых компонент (Independent Computing Architecture, ICA). Данный метод анализирует разделение смешанных сигналов на их исходные источники. Предполагается, что источники независимы друг от друга, а значит не оказывают влияние друг на друга [5].

Попробуем приметь данные методы для классификации текста. В таком случае процесс решение будет состоять из следующих шагов:

1. Загрузка данных для обучения
2. Векторизация текстовых данных
3. Уменьшение размерности полученных векторов
4. Формирование обучающие и тестирующей выборок
5. Подготовка классификатора
6. Обучение классификатора
7. Тестирование классификатора, получение метрик

В роли векторизаторов используем предобученные модели алгоритмов ELMo и BERT. Для ELMo используется модель ruwikiruscorpora_tokens_elmo_1024_2019, на выходе которой получается

эмбеддинг размерностью 1024 [6]. Для BERT – google-bert/bert-base-multilingual-cased, на выходе получается эмбеддинг размерностью 768 [7]. Уменьшаться получаемые эмбеддинги будут до размерностей 700, 500, 250 и 100.

В качестве классификаторов используем нейронную сеть прямого распространения. Количество входов сети соответствует выбранному методу векторизации. Количество выходов равно количеству классов в наборе данных. Нейронная сеть имеет один скрытый слой с 512 нейронами и использует функцию активации ReLU. К выходу модели применяется функция LogSoftmax. Подробная схема модели изображена на рис. 1 (на схеме количество входных нейронов соответствует выходу BERT). В качестве функции потерь используется CrossEntropyLoss, оптимизатора – AdamW. Модель обучалась в течении 100 эпох. Архитектуры была разработана с помощью фреймворка PyTorch [8].

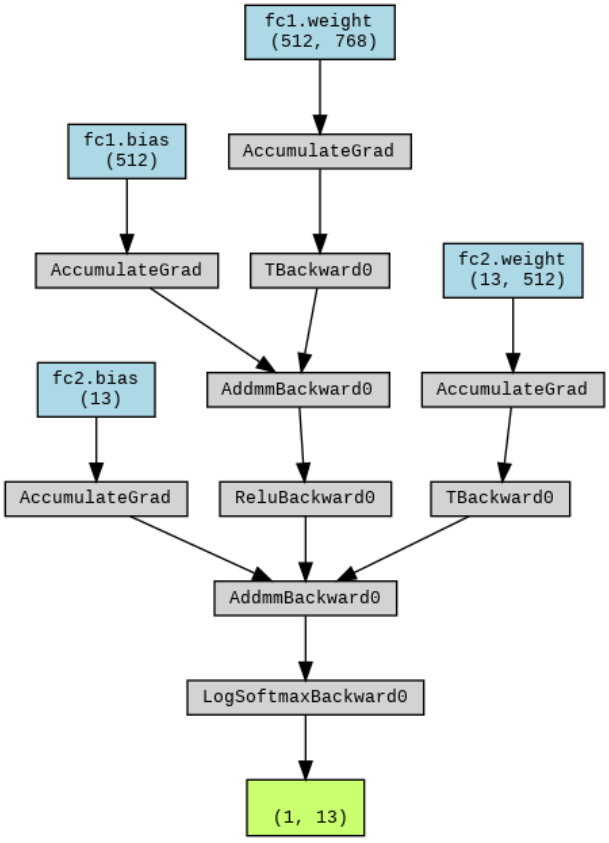

Рис. 1. – Архитектура нейронной сети для классификации

Обучение производилось на наборе данных Russian Social Media Text Classification. В нем содержится 35774 записей, распределенных по 13 классам [9]. На рис. 2 изображено распределение записей по классам в наборе данных.

Рис. 2. – Распределение записей по классам

Результаты обучения моделей представлены в таблице 1. В качестве метрики использовалась взвешенная усредненная F1-мера [10].

Таблица № 1

Результаты обучения классификаторов

Векторизация	**Размерность вектора**					
	1024	768	700	500	250	100
ELMo	0,70	-	-	-	-	-
ELMo + PCA	-	-	0,71	0,71	0,70	0,67
ELMo + ICA	-	-	0,67	0,68	0,68	0,66
BERT	-	0,67	-	-	-	-
BERT + PCA	-	-	0,69	0,69	0,67	0,64
BERT + ICA	-	-	0,64	0,65	0,66	0,63

Изменения размерности векторов незначительно повлияло на результаты обучения классификаторов. Значит, уменьшение размерности словоформенных эмбеддингов можно применять в задачах обработки естественного языка, в которых не требуется их обратное декодирование. К таким задачам относятся классификация, кластеризация, оценка похожести слов. В нашем случае видно, что результаты с использованием метода PCA выше, чем с ICA и без уменьшения размерности. Наилучшие показатели получились при использовании алгоритма векторизации ELMo.

В будущем можно проверить больше методов уменьшения размерности векторов и обучить с их использованием больше различных алгоритмов машинного обучения.

Литература

1. Фролов И. Т. Философский словарь. М. : Республика, 2001. 719 с.

2. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. М. : Финансы и статистика, 1989. 607 с.

3. Сайгин А. А., Федосин С. А. Обзор алгоритмов векторизации текстов на естественном языке // Научно-технический вестник Поволжья. 2024. № 1. С. 99-102.

4. Roweis S. T., Saul L. K. Nonlinear dimensionality reduction by locally linear embedding // science. 2000. V. 290. №. 5500. pp. 2323-2326.

5. Isomura T., Toyoizumi T. A local learning rule for independent component analysis // Scientific reports. 2016. V. 6. №. 1. P. 28073.

6. Kutuzov A., Kuzmenko E. WebVectors: a toolkit for building web interfaces for vector semantic models //Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers 5. – Springer International Publishing, 2017. – pp. 155-161.

7. Devlin J., Chang M. W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. – 2018. URL : arxiv.org/abs/1810.04805 (дата обращения: 01 ноября 2024).

8. PyTorch : сайт – 2016. URL : pytorch.org (дата обращения: 01 ноября 2024).

9. Russian Social Media Text Classification : Kaggle : сайт – 2010. URL : kaggle.com/datasets/mikhailma/russian-social-media-text-classification/data (дата обращения: 01 ноября 2024).

10. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск [пер. с англ. Д. А. Клюшина]. М. : Вильямс, 2011. 520 с.

References

1. Frolov I. T. Filosofskij slovar [Philosophical dictionary]. M. : Respublika, 2001. 719 p.

2. Ajvazyan S. A., Buhshtaber V. M., Enyukov I. S., Meshalkin L. D. Prikladnaya statistika: klassifikaciya i snizhenie razmernosti [Applied statistics: classification and reduction of dimensionality]. M. : Finansy i statistika, 1989. 607 p.

3. Saygin A. A., Fedosin S. A. Nauchno-tekhnicheskij vestnik Povolzhya. 2024. № 1. pp. 99-102.

4. Roweis S. T., Saul L. K. science. 2000. V. 290. №. 5500. pp. 2323-2326.

5. Isomura T., Toyoizumi T. Scientific reports. 2016. V. 6. №. 1. pp. 28073.

6. Kutuzov A., Kuzmenko E. Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers 5. Springer International Publishing, 2017. pp. 155-161.

7. Devlin J., Chang M. W., Lee K., Toutanova K. arXiv preprint arXiv:1810.04805. 2018. URL: arxiv.org/abs/1810.04805 (Date accessed: 01 noyabrya 2024).

8. PyTorch : sajt – 2016. URL: pytorch.org (Date accessed: 01 noyabrya 2024).

9. Russian Social Media Text Classification: Kaggle: sajt – 2010. URL: kaggle.com/datasets/mikhailma/russian-social-media-text-classification/data (Date accessed: 01 noyabrya 2024).

10. Manning K. D., Raghavan P., Shyutce H. Vvedenie v informacionnyj poisk [Introduction to Information Retrieval] [per. s angl. D. A. Klyushina]. M.: Vilyams, 2011. 520 p.

Дата поступления: 3.11.2024

Дата публикации: 1.01.2025