УДК 004.89
DOI: 10.15827/0236-235X.128.650-654
Дата подачи статьи: 13.09.19
2019. Т. 32. № 4. С. 650–654

Особенности применения нейро-сетевых моделей для классификации коротких текстовых сообщений

М.И. Дли [1], *д.т.н., зам. директора по научной работе, midli@mail.ru*
О.В. Булыгина [1], *к.э.н., доцент, baguzova_ov@mail.ru*

[1] *Смоленский филиал Национального исследовательского университета МЭИ, г. Смоленск, 214013, Россия*

В настоящее время органы государственной власти активно развивают технологии электронного взаимодействия с организациями и населением. Одной из ключевых задач в данной сфере является классификация поступающих сообщений, необходимая для их оперативной обработки. Однако особенности таких сообщений (небольшой размер, отсутствие четкой структуры и т.д.) не позволяют применять традиционные подходы к анализу текстовой информации.

Для решения указанной проблемы предложено использовать нейро-сетевые модели (искусственные нейронные сети и нейро-нечеткий классификатор), которые позволяют находить скрытые закономерности в документах, написанных на естественном языке. Выбор конкретного метода определяется подходом к формированию тематических рубрик: сверточные нейронные сети при однозначном определении рубрик, рекуррентные нейронные сети при значимом порядке слов в названиях рубрик, нейро-нечеткий классификатор при пересечении тезаурусов рубрик.

Ключевые слова: *классификация текстов, нейро-нечеткий классификатор, искусственные нейронные сети.*

Активная информатизация всех сфер человеческой деятельности обусловливает развитие компьютерной лингвистики, занимающейся вопросами автоматической обработки текстовой информации. Одной из актуальных задач анализа такой информации является разработка методов классификации документов, написанных на естественном языке. Прежде всего это связано с необходимостью обработки больших объемов электронных сообщений, поступающих на интернет-ресурсы различных организаций и учреждений.

Особенно остро данная проблема стоит перед органами государственной власти, которые активно внедряют технологии электронного взаимодействия с гражданами и организациями. Ежегодный рост объемов обращений, поступающих на интернет-порталы и электронную почту, приводит к необходимости использования систем автоматического анализа обращений с целью оперативного распределения между различными департаментами, которые будут их обрабатывать. В этом случае задача классификации электронных текстовых сообщений сводится к их разнесению по тематическим рубрикам, определяющим направления деятельности различных департаментов.

Вопросам классификации текстов посвящено большое количество публикаций в России и за рубежом. Как показал их анализ, выбор конкретного метода определяется спецификой анализируемых сообщений и особенностями формирования классов (рубрик).

Так, отличительными характеристиками электронных текстовых сообщений, поступающих на интернет-порталы органов государственной власти, являются небольшой размер, отсутствие четкой структуры, свободный стиль изложения, а также разнообразие типов обращений (предложения, заявления, жалобы и т.п.) и рассматриваемых вопросов.

Данные особенности накладывают определенные ограничения на применение традиционных подходов к анализу текстовой информации. В связи с этим целесообразно использовать методы интеллектуального анализа данных, которые позволяют обрабатывать неструктурированные текстовые сообщения в условиях динамичности тезауруса рубрик.

Мощным инсрументом машинного обучения являются искусственные нейронные сети, которые позволяют находить скрытые закономерности в документах, написанных на естественном языке [1, 2]. Например, для решения задач классификации текстов можно использовать несколько топологий нейронных сетей: сверточные [3–6], рекуррентные [7, 8], рекурсивные [9–11] сети и автокодировщики [12]. Кроме того, для классификации текстов могут

применяться нейро-нечеткие модели [13–16], которые позволяют анализировать короткие сообщения в условиях ограниченности статистической информации.

Однако каждый из указанных математических методов имеет свои условия применимости, в значительной степени связанные с особенностями обучения. Поэтому при разработке информационной системы автоматического анализа текстовых сообщений целесообразно использовать несколько моделей, выбор которых определяется постановкой задачи классификации и характеристиками анализируемых текстов.

Принято выделять две группы задач классификации текстов, определяющие, в свою очередь, подход к их решению.

Группа 1. Бинарная классификация, которая отвечает на вопрос, интересно ли пользователю поступившее сообщение (используется, например, при выявлении нежелательной рассылки). Для реализации такой классификации обычно применяется логистическая регрессия.

Группа 2. Мультиклассовая классификация, которая относит сообщение к одному (или нескольким) классу из некоторого множества. Данный тип классификации может быть реализован двумя способами:

– использование функции Softmax, вычисляющей дробные вероятности отнесения сообщения к каждому классу (сумма вероятностей всегда равна 1) и применяемой только в случае, когда сообщение относится к одному классу; функция Softmax часто применяется в искусственных нейронных сетях;

– многократная реализация бинарной классификации, то есть построение отдельного классификатора для каждого класса из всего множества; такой подход может быть реализован с использованием нейро-нечеткого классификатора.

Поставленная в статье научная задача классификации коротких текстовых сообщений, поступающих на интернет-порталы органов государственной власти, является мультиклассовой в силу большого количества рубрик, к которым может быть отнесен запрос.

Критерием выбора способа мультиклассовой классификации является степень связанности рубрик. Так, при наличии пересечения тезаурусов рубрик целесообразно использовать нейро-нечеткий классификатор, в противном случае – искусственные нейронные сети.

Модель классификации текстовых сообщений на основе нейро-нечеткого классификатора

В процессе классификации коротких текстовых сообщений могут возникать ситуации, когда рубрики (классы) не имеют четких границ или их множества пересекаются. В этом случае можно использовать нейро-нечеткий классификатор, объединяющий возможности нечеткой логики и искусственных нейронных сетей.

В общем, нейро-нечеткий классификатор – это тип нейронной сети, которая является адаптивным эквивалентом нечетко-логической модели. Суть данного аппарата заключается в формировании системы нечетких продукционных правил (экспертных знаний о предметной области) и процедуры формирования заключений на основе множества нечетких предпосылок. В этом аппарате алгоритмы нечеткого логического вывода реализованы в виде нейронной сети, имеющей разнородные слои нейронов.

Однако особенности коротких текстовых сообщений, поступающих на интернет-порталы органов государственной власти, не позволяют в явном виде использовать данный математический аппарат, что приводит к необходимости его модификации.

В известных моделях, реализованных в форме нейро-нечеткого классификатора, текстовое сообщение представляется в виде массива бинарных значений, характеризующих наличие в нем слов из тезауруса каждой рубрики. Однако этот подход сложно реализовать в условиях динамичности рубрик из-за необходимости перестроения нейро-нечеткой модели. Для решения данной задачи можно использовать каскадную модель классификации текстовых сообщений, включающую:

– подмодель предварительного анализа (выявление в сообщении значимых слов и формирование множества синтаксических групп);

– подмодель формализации (оценка степени принадлежности синтаксических групп к выделенным рубрикам);

– множество подмоделей оценки принадлежности сообщения к каждой рубрике (в виде нейро-нечеткого классификатора);

– подмодель выбора наиболее подходящей рубрики.

Такой подход к классификации текстовых сообщений позволяет анализировать документы небольшого размера на основе их унифицированного представления.

Классификация текстовых сообщений с использованием искусственных нейронных сетей

Анализ публикаций по классификации текстов с использованием искусственных нейронных сетей выявил, что хорошие результаты показывают сверточные сети, которые способны обрабатывать короткие сообщения.

Сверточные сети – это тип нейронных сетей прямого распространения, когда сигнал идет последовательно по нейронам (от первого слоя к последнему). Они представляют собой чередование сверточных, субдискретизирующих и полносвязных слоев на выходе.

Изначально сверточные сети разрабатывались для анализа изображений. Хорошие результаты в данной области способствовали их применению для решения других классификационных задач, в том числе для анализа текстовых сообщений.

Данный тип искуственных нейронных сетей предлагается использовать для классификации текстовых сообщений, когда рубрики не пересекаются. В этом случае на вход сверточной сети подается короткое сообщение, в котором каждое слово определяется вектором фиксированной длины (например, может использоваться алгоритм word2vec). Для выходного слоя целессообразно использовать функцию Softmax, реализующую мультиклассовую классификацию.

Однако нередко возникают ситуации, когда при задании тематических рубрик важен порядок слов в названии и словосочетаниях, определяющих их тезаурус. В этом случае возникает задача классификации последовательностей. Одним из успешных инструментов для решения данной задачи являются рекуррентные сети.

Рекуррентные сети – это тип нейронных сетей с обратными связями, когда нейроны используют информацию предыдущего слоя и данные о состоянии этих нейронов на предшествующем проходе. Аналогичным образом на вход сети текстовое сообщение подается в виде вектора, а на выходе используется функция Softmax.

Применение моделей классификации коротких текстовых сообщений, построенных на основе искусственных нейронных сетей, возможно только после предварительной обработки (токенизация, морфологический анализ, удаление стоп-слов и т.п.) и индексации текстового сообщения (построение числовой модели).

Другим условием применения искусственных нейронных сетей является формирование большого объема данных примеров, используемых для обучения моделей.

На основе вышесказанного была предложена обощенная процедура проведения классификации коротких текстовых сообщений, поступающих на интернет-порталы органов государственной власти, которая представлена на рисунке.

Обощенная процедура классификации коротких текстовых сообщений

A generalized procedure for the classification of short text messages

Таким образом, в статье предложен новый подход к использованию нейро-сетевых алгоритмов для решения задач классификации электронных неструктурированных текстовых документов, поступающих на интернет-порталы органов государственной власти. Выделены три типа ситуаций, для которых могут использоваться следующие нейро-сетевые алгоритмы: сверточные нейронные сети в случае однозначного определения тематических рубрик, рекуррентные нейронные сети, когда важен порядок слов при определении рубрик и значимых слов, нейро-нечеткий классификатор при пересечении тезаурусов рубрик.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-01-00558.

Литература

1. Bengio Y., Ducharme R., Vincent P., Jauvin C. A neural probabilistic language model. J. of Machine Learning Research, 2003, vol. 3, pp. 1137–1155.
2. Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P. Natural language processing (almost) from scratch. J. of Machine Learning Research, 2011, vol. 12, pp. 2493–2537.
3. Kim Y. Convolutional neural networks for sentence classification. Proc. IEMNLP, 2014, pp. 1746–1751.
4. Krizhevsky A., Sutskever I., Hinton G. Imagenet classification with deep convolutional neural networks. Proc. NIPS, 2012, pp. 1106–1114. DOI: 10.1145/3065386.
5. LeCun Y. Text understanding from scratch. Computer Science Department, 2016. URL: https://arxiv.org/pdf/1502.01710.pdf (дата обращения: 10.09.2019).
6. Zhang X., Zhao J., LeCun Y. Character-level convolutional networks for text classification. Proc. NIPS, 2015, pp. 649–657.
7. Kalchbrenner N., Blunsom P. Recurrent convolutional neural networks for discourse compositionality. Workshop CVSC, 2013, pp. 119–126.
8. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality. Proc. NIPS, 2013, pp. 3111–3119.
9. Iyyer M., Enns P., Boyd-Graber J., Resnik P. Political ideology detection using recursive neural networks. Proc. ACL, 2014, pp. 1113–1122. DOI: 10.3115/v1/P14-1105.
10. Socher R., Huang E., Pennington J., Ng A., Manning C. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Proc. NIPS, 2011, vol. 24, pp. 801–809.
11. Socher R., Perelygin A., Wu J., Chuang J., Manning C., Ng A., Potts C. Recursive deep models for semantic compositionality over a sentiment Treebank. Proc. EMNLP, 2013, pp. 1631–1642.
12. Socher R., Pennington J., Huang E., Ng A., Manning C. Semi-supervised recursive autoencoders for predicting sentiment distributions. Proc. EMNLP, 2011, pp. 151–161.
13. Dli M., Bulygina O., Kozlov P., Ross G. Developing the economic information system for automated analysis of unstructured text documents. J. of Applied Informatics, 2018, vol. 13, no. 5, pp. 51–57.
14. Dli M., Bulygina O., Kozlov P. Development of multimethod approach to rubrication of unstructed electronic text documents in various conditions. Proc. Intern. RusAutoCon., Sochi, 2018, pp. 1–5. DOI: 10.1109/RUSAUTOCON.2018.8501815.
15. Tukaev D., Bulygina O., Kozlov P., Morozov A., Chernovalova M. Cascade neural-fuzzy model of analysis of short electronic unstructured text documents using expert information. ARPN JEAS, 2018, vol. 13, no. 21, pp. 8531–8536.
16. Avdeenko T.V., Makarova E.S. Acquisition of knowledge in the form of fuzzy rules for cases classification. LNC, 2017, vol. 10387, pp. 536–544. DOI: 10.1007/978-3-319-61845-6_53.

Software & Systems
DOI: 10.15827/0236-235X.128.650-654

Received 13.09.19
2019, vol. 32, no. 4, pp. 650–654

Features of using neural network models to classify short text messages

M.I. Dli [1], *Dr.Sc. (Engineering), Deputy Director on Scientific Work, midli@mail.ru*
O.V. Bulygina [1], *Ph.D. (Economics), Associate Professor, baguzova_ov@mail.ru*

[1] *Smolensk Branch of the Moscow Power Engineering Institute, Smolensk, 214013, Russian Federation*

Abstract. Nowadays, public authorities are actively developing technologies of electronic interaction with organizations and citizens. One of the key tasks in this area is classification of incoming messages for their

operational processing. However, the features of such messages (small size, lack of a clear structure, etc.) do not allow using traditional approaches to the analysis of textual information.

To solve this problem, it is proposed to use neural network models (artificial neural networks and neuro-fuzzy classifier), which allow finding hidden patterns in documents written in a natural language. The choice of a specific method is determined by the approach to forming thematic headings: convolutional neural networks (for unambiguous definition of rubrics); recurrent neural networks (for significant word order in the title of rubrics); neuro-fuzzy classifier (for intersecting thesauri of rubrics).

Keywords: text classification, neuro-fuzzy classifier, artificial neural networks.

Acknowledgements. *The research was financially supported by RFBR within the framework of a research project No. 18-01-00558.*

References

1. Bengio Y., Ducharme R., Vincent P., Jauvin C. A Neural probabilistic language model. *J. of Machine Learning Research*. 2003, vol. 3, pp. 1137–1155.
2. Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P. Natural language processing (almost) from scratch. *J. of Machine Learning Research*. 2011, vol. 12, pp. 2493–2537.
3. Kim Y. Convolutional neural networks for sentence classification. *IEMNLP*. 2014, pp. 1746–1751.
4. Krizhevsky A., Sutskever I., Hinton G. Imagenet classification with deep convolutional neural networks. *Proc. NIPS'12*. 2012, pp. 1106–1114. DOI: 10.1145/3065386.
5. LeCun Y. *Text Understanding From Scratch*. Computer Science Department, 2016. Available at: https://arxiv.org/pdf/1502.01710.pdf (accessed September 10, 2019).
6. Zhang X., Zhao J., LeCun Y. Character-level convolutional networks for text classification. *NIPS Proc*. 2015, pp. 649–657.
7. Kalchbrenner N., Blunsom P. Recurrent convolutional neural networks for discourse compositionality. *Workshop on CVSC*. 2013, pp. 119–126.
8. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality. *Proc. of NIPS'13*. 2013, pp. 3111–3119.
9. Iyyer M., Enns P., Boyd-Graber J., Resnik P. Political ideology detection using recursive neural networks. *Proc. of ACL*. 2014, pp. 1113–1122. DOI: 10.3115/v1/P14-1105.
10. Socher R., Huang E., Pennington J., Ng A., Manning C. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Proc. of NIPS'11*. 2011, vol. 24, pp. 801–809.
11. Socher R., Perelygin A., Wu J., Chuang J., Manning C., Ng A., Potts C. Recursive deep models for semantic compositionality over a sentiment Treebank. *Proc. EMNLP*. 2013, pp. 1631–1642.
12. Socher R., Pennington J., Huang E., Ng A., Manning C. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proc. EMNLP*. 2011, pp. 151–161.
13. Dli M., Bulygina O., Kozlov P., Ross G. Developing the economic information system for automated analysis of unstructured text documents. *J. Applied Informatics*. 2018, vol. 13, no. 5, pp. 51–57.
14. Dli M., Bulygina O., Kozlov P. Development of multimethod approach to rubrication of unstructed electronic text documents in various conditions. *Proc. Intern. Russ. Automation Conf*. Sochi, 2018. DOI: 10.1109/RUSAUTOCON.2018.8501815.
15. Tukaev D., Bulygina O., Kozlov P., Morozov A., Chernovalova M. Cascade neural-fuzzy model of analysis of short electronic unstructured text documents using expert information. *ARPN JEAS*. 2018, vol. 13, no. 21, pp. 8531–8536.
16. Avdeenko T.V., Makarova E.S. Acquisition of knowledge in the form of fuzzy rules for cases classification. *LNCS*. 2017, vol. 10387, pp. 536–544. DOI: 10.1007/978-3-319-61845-6_53.

Для цитирования

Дли М.И., Булыгина О.В. Особенности применения нейро-сетевых моделей для классификации коротких текстовых сообщений // Программные продукты и системы. 2019. Т. 32. № 4. С. 650–654. DOI: 10.15827/0236-235X.128.650-654.

For citation

Dli M.I., Bulygina O.V. Features of using neural network models to classify short text messages. *Software & Systems*. 2019, vol. 32, no. 4, pp. 650–654 (in Russ.). DOI: 10.15827/0236-235X.128.650-654.