ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА "МИФИ", 2020, том 9, № 5, с. 475–480

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.032.26

ПРИМЕНЕНИЕ ОГРАНИЧЕННОЙ МАШИНЫ БОЛЬЦМАНА ДЛЯ РЕШЕНИЯ ЗАДАЧИ АВТОРСКОГО ПРОФИЛИРОВАНИЯ РУССКОЯЗЫЧНЫХ ТЕКСТОВ[1]

© 2020 г. А. Г. Сбоев[1,2,*], Р. Б. Рыбка[1], Ю. А. Давыдов[1], А. А. Селиванов[1]

[1] НИЦ "Курчатовский институт", Москва, 123182, Россия

[2] Национальный исследовательский ядерный университет "МИФИ", Москва, 115409, Россия

**e-mail: sag111@mail.ru*

Поступила в редакцию 01.10.2020 г.
После доработки 03.11.2020 г.
Принята к публикации 24.11.2020 г.

В данной работе исследуется возможность применения ограниченной машины Больцмана (ОМБ) для решения задачи авторского профилирования текстов на русском языке на примере определения пола и возраста автора. ОМБ используется в качестве трансформера, извлекающего полезные признаки из документов, слова в которых закодированы при помощи морфологических тегов. Классификация осуществляется при помощи составного двухслойного модуля, включающего в себя MultinomialNB и LinearSVC. В рамках поставленной задачи используются четыре корпуса документов, три из которых размечены по полу автора (в том числе с его имитацией), четвертый же – по возрасту. Проведенные эксперименты показывают, что построенная модель успешно решает поставленные перед ней задачи, превосходя baseline-модель (LinearSVC) в среднем (по всем четырем корпусам) на 7.5% по f1-score. Также представляется сравнение полученных результатов с результатами других моделей из литературных источников (в частности, с использованием сложной модели на основе комбинации сверточной нейронной сети и LSTM), демонстрирующее эффективность созданной модели и ее устойчивость по набору представленных корпусов.

DOI: 10.1134/S2304487X20050144

1. ВВЕДЕНИЕ

Ограниченная машина Больцмана (ОМБ) – разновидность стохастической (основанной на принципах статистической физики, в том числе на применении канонического распределения Гиббса) нейронной сети, которая определяет распределение вероятности на входных образцах данных [1]. ОМБ активно применяются как для извлечения признаков из изображений, документов и др. (см., например, [2–6]), так и непосредственно для классификации (такая модификация ОМБ обучается в supervised-манере, см. [7, 8]).

В настоящей работе исследуется возможность применения ОМБ для решения задачи авторского профилирования текстов на русском языке на примере определения пола и возраста автора. В работе [8] эта задача была решена для англоязычных текстов при помощи классификационной ОМБ. Авторы использовали для кодирования документов частотно-словарный подход. Его привлекательная сторона заключается в возможности анализировать текст без какой бы то ни было глубокой предобработки. Однако в общем случае, как указывается в литературе, этот подход требует существенно большего набора обучающих примеров. Так, например, корпус PAN-AP-13, использованный авторами статьи [8], содержит более 400000 примеров. В то же время, доступные русскоязычные корпуса, подобранные для решения задачи авторского профилирования, значительно меньше по объему (массивы документов, используемые в данной работе, меньше PAN-AP-13 на два порядка).

Исходя из этих ограничений авторами настоящей работы был выбран иной подход: ОМБ здесь используется для извлечения полезных признаков из документов, закодированных при помощи морфологических тегов. Важно также отметить, что такой выбор кодирования, будучи несколько более трудоемким с точки зрения предобработки корпусов, не зависит от выбора словаря, по которому производится частотно-словарное кодирование, а значит, является более универсальным.

[1] Работа была выполнена с использованием оборудования центра коллективного пользования "Комплекс моделирования и обработки данных исследовательских установок мега-класса" НИЦ "Курчатовский институт", http://ckp.nrcki.ru/.

2. КОРПУСА ДАННЫХ

Для изучения возможностей, представляемых ОМБ в контексте решения задачи авторского профилирования, было использовано четыре корпуса данных: первый из них (RusPersonality (RusPer)) состоит из 1150 сочинений, написанных студентами и в дальнейшем размеченных экспертами, второй, третий и 4-й (Gender Imitation crowdsource “a” (GI cs “a”), GI cs “ab” и Age Imitation crowdsource “a” (AI cs “a”)) – из 1664, 3330 и 1680 документов соответственно, собранных при помощи краудсорсинговой платформы. При этом в третьем корпусе содержатся тексты с имитацией: авторам было дано задание сымитировать стиль человека противоположного пола. Четвертый же корпус, в отличие от трех других, разбит на три класса по возрасту автора (20–30 лет, 30–40 лет, 40–50 лет). Все четыре корпуса были предварительно сбалансированы по классам. Вычислительные эксперименты проводились на основе стратифицированной кросс-валидации со смешением (stratified shuffle split) с пятью разбиениями корпуса. В каждом разбиении 80% корпуса представляло тренировочное множества, 20% – тестовое. После этого в каждом разбиении 10% тренировочного множества использовалось для валидации модели, таким образом конечное соотношение множеств для каждой итерации кросс-валидации:

- 72% – тренировочное множество;
- 8% – валидационное множество;
- 20% – тестировочное множество.

Также обязательным условием в процессе разбиения было, чтобы в тренировочном и тестировочном множествах были документы, написанные разными авторами. В результате для каждого корпуса было получено пять фолдов, пригодных для проведения кросс-валидации. Далее документы были векторизованы: каждое слово было закодировано вектором, содержащим морфологические признаки (часть речи, падеж и т.д.) длиной 58. Этот способ представления данных уже доказал свою эффективность в задаче авторского профилирования (см. [9]).

3. ТОПОЛОГИЯ СЕТИ И АЛГОРИТМЫ

В данной работе используется реализация классической ОМБ из библиотеки scikit-learn [10]: это так называемая BernoulliRBM, обучаемая при помощи алгоритма PCD (Persistent Contrastive Divergence; также встречается название “стохастическая максимизация правдоподобия”). Ее важной особенностью является работа только с бинарным входом, т.е. обучающие примеры должны состоять только из нулей и единиц.

Как уже было сказано, для решения задачи определения пола или возраста автора была использована связка из ОМБ и следующего за ней классификатора. То, какой именно классифицирующий алгоритм использовать – очень существенный вопрос, ответ на который был получен с использованием библиотеки TPOT [11], позволяющей по заданному множеству обучающих примеров подобрать оптимальный классификатор из набора, предоставляемого scikit-learn. Закодированные морфологическими признаками слова из корпуса RusPersonality объединялись в триграммы длиной 174 (со сдвигом 2) и подавались в ОМБ с числом нейронов на видимом слое, равном входной размерности примеров, и с величиной скрытого слоя, равной 600 нейронов (тестировались и иные конфигурации, в том числе сжимающие; выяснилось, что расширяющая ОМБ дает лучшие результаты, причем чем больше скрытый слой (до определенного порога), тем лучше). После документы, входящие в корпус, по очереди преобразовывались обученной ОМБ: фактически, каждая триграмма в документе кодировалась логистическими функциями активации, связанными со скрытым слоем ОМБ. Далее для каждого документа триграммы усреднялись – в итоге для документов были получены вектора, массив которых, соответствующий тренировочному множеству, подавался в TPOT, который обучался до тех пор, пока внутренняя кросс-валидационная точность не переставала меняться в течение десяти эпох. Полученный классификатор состоит из двух частей: первая из них (**MultinomialNB**) выступает в роли эстиматора, вторая (**LinearSVC**) – осуществляет классификацию непосредственно.

TPOT не дает ответа на вопрос, оптимально ли была выбрана топология ОМБ, поэтому необходимо дополнительно оптимизировать набор гиперпараметров ОМБ (число нейронов на скрытом слое, learning rate и т.д.) совместно с набором гиперпараметров классификатора.

Для этой цели был использован HyperOpt [12] (библиотека, осуществляющая подбор оптимальных гиперпараметров сети исходя из заданного тренировочного множества), запуск которого осуществлялся на одном из фолдов, соответствующих корпусу Toloka_a (предварительные тесты показали, что он анализируется несколько хуже, а значит, подбирая оптимальную конфигурацию для него, можно ожидать, что более простой корпус RusPersonality также будет классифицироваться успешно). Финальная топология целой модели представлена на рис. 1. Параметры, использованные в экспериментах:

- **RBM**: n_visible = 58 * 3 – число нейронов в видимом слое; n_hidden = 560 – число нейронов в скрытом слое;
- **MultinomialNB**: alpha = 0.001; fit_prior = True;

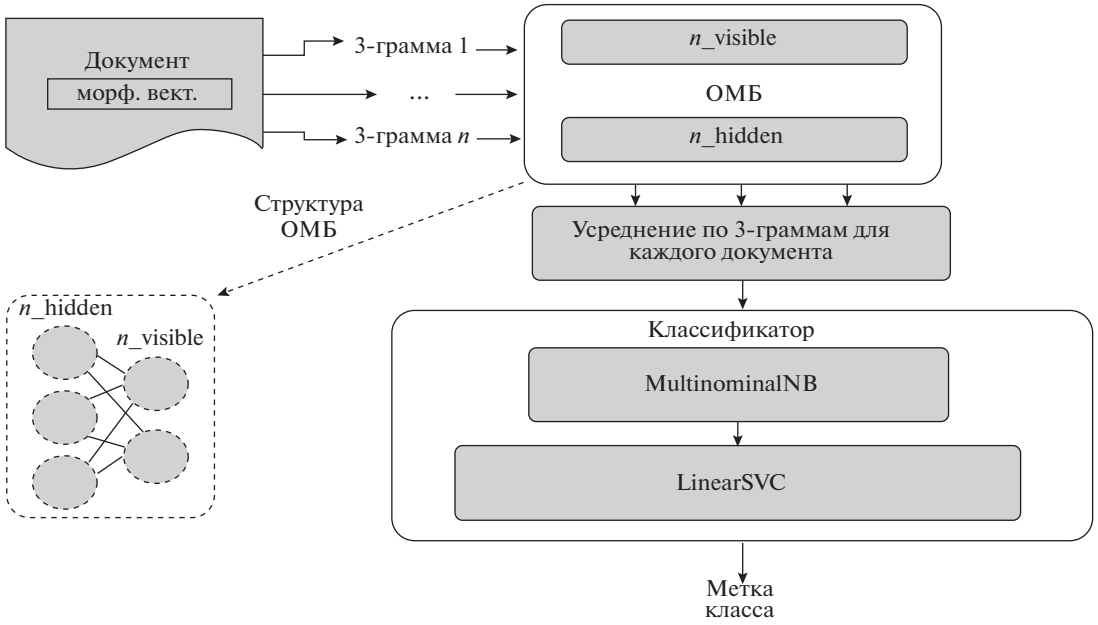

Рис. 1. Общая топология сети RBM_SVC.

• **LinearSVC**: C = 10.0; dual = False; loss = = “squared_hinge”; penalty = “l2”; tol = 0.0001; max_iter = 100000.

4. ЭКСПЕРИМЕНТЫ

Чтобы установить, насколько хорошо связка ОМБ+SVC подходит для решения задачи авторского профилирования, модель была обучена и протестирована на всех трех корпусах следующим образом: на каждом из пяти фолдов производилось по одному запуску (без фиксации random state), после чего для полученного MacroF1-score вычислялось выборочное среднее по всем пяти фолдам.

Чтобы оценить уровень полученных результатов, необходимо определить нижнюю границу точности, относительно которой и будет производиться оценка. В данном случае для всех корпусов можно построить похожую модель, не включающую в себя ОМБ, и провести на ней серию экспериментов, в остальном полностью идентичную описанной выше: слова, входящие в состав документов, кодируются при помощи морфологических признаков, объединяются в триграммы с шагом 1, после чего усредняются и подаются в классификатор.

Исходя из этого, была построена baseline-модель, состоящая из LinearSVC с теми же параметрами, что и у LinearSVC, входящего в модель RBM_SVC. Результаты экспериментов в сравнении с доступными историческими данными[2] приведены в табл. 1–4.

5. ОБСУЖДЕНИЕ

Из приведенных выше экспериментальных результатов можно заключить, что составная нейронная сеть из ОМБ и двухслойного классификатора на основе LinearSVC эффективно решает задачу авторского профилирования документов на русском языке. Прирост macrof1-score по сравнению с baseline-моделью (LinearSVC) в среднем по всем четырем использованным корпусам составляет 7.5%, достигая 8% на корпусе RusPer.

Для дополнительной оценки полученных точностей можно обратиться к результатам, полученным нами на других моделях. Так, в сравнении с работой [13] для сбалансированного корпуса RusPer, закодированного морфологическими признаками, модель, построенная на основе сверточной нейронной сети и слоя LSTM, показывает macro f1-score, равный 0.81+/–0.04. В пределах погрешности этот результат довольно близок к полученному в данной работе (0.77+/–0.02). Следует также отметить, что модель под кодовым именем ModelNN частично учитывает структуру

[2] Исторические данные были получены следующим образом: в случае ModelNN из статьи [13] была взята топология сети и протестирована на том же разбиении, на котором проводились эксперименты с RBM_SVC и LinearSVC из данной работы. Аналогичным образом для AI cs “a” был пересчитан TF-IDF+LinearSVC (сокращенно TFIDF) из статьи [14].

Таблица 1. Результаты экспериментов на корпусе RusPer

RusPer	fold 0	fold 1	fold 2	fold 3	fold 4	AVG	STD
RBM_SVC	0.77	0.78	0.83	0.71	0.76	0.77	0.02
LinearSVC	0.64	0.69	0.73	0.68	0.73	0.69	0.03
ModelNN [13]						0.81	0.04

Таблица 2. Результаты экспериментов на корпусе GIcs “a”

GI cs “a”	fold 0	fold 1	fold 2	fold 3	fold 4	AVG	STD
RBM_SVC	0.65	0.71	0.70	0.69	0.68	0.69	0.02
LinearSVC	0.55	0.59	0.62	0.58	0.59	0.59	0.02
ModelNN [13]						0.77	0.04

Таблица 3. Результаты экспериментов на корпусе GIcs “ab”

GI cs “ab”	fold 0	fold 1	fold 2	fold 3	fold 4	AVG	STD
RBM_SVC	0.57	0.55	0.58	0.54	0.56	0.56	0.02
LinearSVC	0.56	0.54	0.59	0.55	0.56	0.56	0.02
ModelNN [13]						0.51	0.02

Таблица 4. Результаты экспериментов на корпусе AIcs “a”

AI cs “a”	fold 0	fold 1	fold 2	fold 3	fold 4	AVG	STD
RBM_SVC	0.57	0.55	0.58	0.54	0.56	0.56	0.02
LinearSVC	0.47	0.47	0.49	0.50	0.48	0.48	0.01
TFIDF [14]	0.56	0.55	0.57	0.56	0.52	0.55	0.01
ModelNN [13]	0.49	0.47	0.47	0.44	0.43	0.46	0.02

текста на уровне последовательностей, поэтому является хорошим baseline для сравнения. Результат на корпусе Toloka_a демонстрирует некоторое ее превосходство по отношению к результату, полученному на RBM_SVC (0.02 над статистической погрешностью). Для корпуса с имитацией пола можно констатировать более высокую эффективность RBM_SVC модели. То же самое можно сказать и о задаче определения возрастной группы. Таким образом, можно заключить, что предлагаемая модель демонстрирует хорошую точность и устойчивость на ряде корпусов, что является ее сильной стороной.

6. ВЫВОДЫ

В целом можно заключить, что предложенный в данной работе метод использования ОМБ для решения задачи авторского профилирования оказывается достаточно перспективным. Построенная модель превосходит baseline-модель (LinearSVC) в среднем (по всем четырем корпусам) на 7.5% по f1-score, что доказывает эффективность ОМБ как экстрактора полезных признаков, а также ее устойчивость по набору представленных корпусов. Сказанное дает основания полагать, что дальнейшие исследования в рамках этого подхода позволят добиться дальнейшего прироста точности классификации, а также развить на его основе классифицирующие модели для анализа русскоязычных текстов.

7. БЛАГОДАРНОСТИ

Работа выполнена при поддержке гранта РФФИ № 18-29-10084 мк.

СПИСОК ЛИТЕРАТУРЫ

1. *Goodfellow I., Yoshua Bengio, Courville A.* Deep Learning. MIT Press, 2016.
2. *Hinton G.E., Osindero S., Yee Whye Teh.* A fast learning algorithm for deep belief nets // Neural Computation. 2006. V. 18. P. 1527–1554.
3. *Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle.* Greedy layer-wise training of deep networks / In: Advances in Neural Information Processing Systems 19 (NIPS’06), Ed. by *B. Schölkopf, J. Platt, T. Hoffman*. 2007. MIT Press, 2007. P. 153–160.
4. *Welling M., Sutton Ch.* Learning in Markov random fields with contrastive free energies / In: Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS’05), 2005. P. 397–404.
5. *Salakhutdinov R., Hinton G.E.* Semantic hashing / In: Proceedings of the 2007 Workshop on Information Retrieval and applications of Graphical Models (SIGIR’07), Amsterdam: Elsevier, 2007.
6. *Mnih A., Hinton G.E.* Three new graphical models for statistical language modelling / In: Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML’07). Ed. by Zoubin Ghahramani, ACM, 2007. P. 641–648.
7. *Larochelle H. et al.* Learning Algorithms for the Classification Restricted Boltzmann Machine // J. Mach. Learn. Res. 2012. V. 13. P. 643–669.
8. *Antkiewicz M., Kuta Marcin, Kitowski J.* Author Profiling with Classification Restricted Boltzmann Machines. 2017.
9. *Sboev A., Litvinova T., Gudovskikh D., Rybka R., Moloshnikov I.* Machine Learning Models of Text Categorization by Author Gender Using Topic-independent Features // Procedia Computer Science, 2016. V. 101. P. 135–142.
10. *Pedregosa et al.* Scikit-learn: Machine Learning in Python. JMLR, 2011. V. 12. P. 2825–2830.
11. *Trang T. Le, Weixuan Fu, Jason H. Moore.* Scaling tree-based automated machine learning to biomedical big data with a feature set selector // Bioinformatics. 2020. V. 36. № 1. P. 250–256.

12. *Bergstra J., Yamins D., Cox D.D.* Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures / To appear in Proc. of the 30th International Conference on Machine Learning (ICML 2013). 2013.
13. *Sboev A., Gudovskikh D., Moloshnikov I., Rybka R.* A gender identification of text author in mixture of Russian multi-genre texts with distortions on base of data-driven approach using machine learning models // AIP Conference Proceedings. 2019. V. 2116. P. 270006. https://doi.org/10.1063/1.5114280
14. *Sboev A., Rybka R., Moloshnikov I., Gudovskikh D., Litvinova T.* To the question of data-driven identification of author's age for Russian texts with age deceptions using machine learning // Journal of Physics: Conf. Ser. 2019. V. 1205. P. 012049.

Vestnik Nacional'nogo Issledovatel'skogo Yadernogo Universiteta "MIFI", 2020, vol. 9, no. 5, pp. 475–480

Application of Restricted Boltzmann Machines to Solve the Problem of Author's Profiling of Russian Texts

A. G. Sboev[a,b,#], R. B. Rybka[a], Yu. A. Davydov[a], and A. A. Selivanov[a]

[a] *National Research Center Kurchatov Institute, Moscow, 123182 Russia*
[b] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409 Russia*
[#] *e-mail: sag111@mail.ru*

Received October 1, 2020; revised November 3, 2020; accepted November 24, 2020

Abstract—The possibility of using the restricted Boltzmann machine (RBM) to solve the problem of author's profiling of Russian texts has been studied on the example of determining the gender and age of an author. The restricted Boltzmann machine is used as a transformer that extracts useful features from documents, where words are encoded using morphological tags. The classification is carried out using a composite two-layer module, which includes MultinomialNB and LinearSVC. Within this task, four corpuses of documents are used, three of which are classified by the gender of the author, and the fourth one, by age. The experiments show that the constructed model successfully solves the tasks assigned to it, surpassing the baseline model (LinearSVC) on average (for all four corpuses) by 7.5% in terms of f1-score. In addition, of the results are compared with the results of other models from the literature (in particular, using a complex model, based on a convolutional neural network and LSTM). This comparison shows the efficiency of the constructed composite neural network based on the RBM and the stability of its results on a set of presented corpora.

Keywords: artificial neural networks, natural language processing, text classification, author profiling, restricted Boltzmann machine, RBM, energy-based neural networks

DOI: 10.1134/S2304487X20050144

REFERENCES

1. Goodfellow I., Yoshua Bengio, Courville A., *Deep Learning,* MIT Press, 2016.
2. Hinton G.E., Osindero S., Yee Whye Teh, A fast learning algorithm for deep belief nets, *Neural Computation,* 2006, vol. 18, pp. 1527–1554.
3. Yoshua Bengio, Lamblin P., Popovici D., Larochelle H., Greedy layer-wise training of deep networks, in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, Schölkopf, B., Platt J., Hoffman T., Ed., MIT Press, 2007, pp. 153–160.
4. Welling M., Sutton Ch., Learning in Markov random fields with contrastive free energies, in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS'05),* 2005, pp. 397–404.
5. Salakhutdinov R., Hinton G.E., Semantic hashing, in *Proceedings of the 2007 Workshop on Information Retrieval and Applications of Graphical Models (SIGIR'07)*, Amsterdam: Elsevier, 2007.
6. Mnih A., Hinton G.E., Three new graphical models for statistical language modeling, in *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07),* Zoubin Ghahramani, Ed., ACM, 2007, pp. 641–648.
7. Larochelle H. et al., Learning Algorithms for the Classification Restricted Boltzmann Machine, *J. Mach. Learn. Res.*, 2012, vol. 13, pp. 643–669.
8. Antkiewicz M., Kuta Marcin, Kitowski J., Author Profiling with Classification Restricted Boltzmann Machines. 2017.
9. Sboev A., Litvinova T., Gudovskikh D., Rybka R., Moloshnikov I., Machine Learning Models of Text Categorization by Author Gender Using Topic-independent Features, *Procedia Computer Science*, 2016, vol. 101, pp. 135–142.

10. Pedregosa, et al., *Scikit-learn: Machine Learning in Python*, JMLR, 2011, vol. 12, pp. 2825–2830.

11. Trang T.Le, Weixuan Fu, Jason H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector, *Bioinformatics*, 2020, vol. 36, no. 1, pp. 250–256.

12. Bergstra J., Yamins D., Cox D.D., Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures, *To appear in Proc. of the 30th International Conference on Machine Learning (ICML 2013),* 2013.

13. Sboev A., Gudovskikh D., Moloshnikov I., Rybka R., A gender identification of text author in mixture of Russian multi-genre texts with distortions on base of data-driven approach using machine learning models, *AIP Conference Proceedings*, 2019, vol. 2116, p. 270006. 10.1063/1.5114280.

14. Sboev A., Rybka R., Moloshnikov I., Gudovskikh D., Litvinova T., To the question of data-driven identification of author's age for Russian texts with age deceptions using machine learning, *J. Phys.: Conf. Ser.,* 2019, vol. 1205, p. 012049.