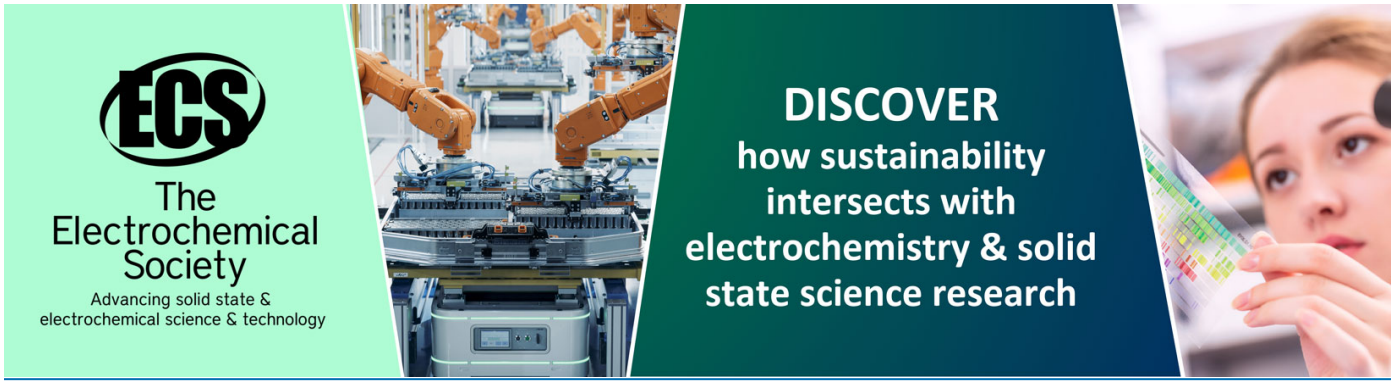
A Text classification algorithm based on topic model and convolutional neural network



A Text classification algorithm based on topic model and convolutional neural network

Junwei Ge[1], Songce Lin[1*], Yiqiu Fang[1]

[1]College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

[*]584116629@qq.com

Abstract. Based on the neural topic model ProdLDA and convolutional neural network, this paper proposes a text classification algorithm based on topic model and convolutional neural network. Firstly, the text information is modeled on the word vector model, then the convolutional neural network is used to extract the granularity features of high-dimensional text, and the neural topic model ProdLDA is used to extract the potential topic features. Then, the connection layer is established to connect the text features, and finally the classification layer is processed. At the same time, a new topic feature introduction method is used in the process of extracting topic features. Experimental results show that this algorithm can effectively improve the performance of text classification.

1. Introduction

With the increasing scale of Internet, People have more and more access to information, and the information they receive becomes more diverse, such as news media, entertainment information, social user messages, short video messages, and so on. In the face of more diversified information on the Internet, people need more efficient information classification processing. Therefore, how to classify the text information accurately and quickly has become a hot topic in current research.

In the research of text classification, the traditional text classification model such as vector space model (VSM) [1], through the calculation between vectors to process the text information, this classification model is prone to dimensional disaster, loss of semantic information and other issues. After that, the researchers proposed a text classification algorithm based on topic model by introducing statistical model knowledge. The origin of topic model is latent semantic analysis LSA (La-tent Semantica Analysisi)[2]. Through singular value decomposition, the high-dimensional document vector is approximately mapped to a low-dimensional potential semantic space, so as to reduce the document dimension and eliminate the synonymy and polysemy of words. Subsequently, Blei[3] proposed the Latent Dirichlet Allocation model (LDA) in 2003. It is a three-level Bayesian model, which can extract potential topic models from corpus, provide an effective way to quantify research topics, and is widely used in text classification. In order to better adapt to the problem of sparse text classification data, Cheng X[4] ignored the concept of document, and proposed Biterm Topic Model (BTM), which transformed the words in the document into word pairs, and directly modeled the word pair set, so that the model had more abundant word co-occurrence information and alleviated the sparsity problem of short text. The essence of topic model is an unsupervised machine learning model. The high-dimensional text word space is represented as a low-dimensional topic space. At the same time, the potential topic features are calculated by using probability model, and the topic space is used for calculation and classification.

These models and their variants have been widely used in various fields of text classification and achieved considerable results. Subsequently, researchers improved the topic model by combining machine learning, and Zhibo Wang[5] supplemented the context semantic information.

In recent years, scholars have mainly completed text feature extraction through deep learning. In 2014, Yoon Kim [6] proposed a text CNN model, which is a kind of convolutional neural network (CNN). The application of CNN in text classification uses multiple convolutional layers of different sizes to extract the key information in sentences, so as to better capture the local relevance and realize the text feature extraction. Then, the recurrent neural network (RNN) for sequence processing is widely used in natural language processing, but there are problems of gradient disappearance and gradient explosion. In order to solve this problem, the long short term memory (LSTM) [8] is designed by improving the network structure of RNN, which is widely used in long text classification. Compared with traditional text classification algorithms, CNN and other deep learning algorithms significantly improve the classification effect and efficiency.

With the gradual application of deep learning to natural language processing, in recent years, many scholars have proposed topic models combined with deep learning ideas, and researchers have begun to propose topic models based on neural networks. This method mainly uses neural networks, such as feedforward neural network and variational self coding network, to reconstruct the text generation process of topic model, and add sparse constraints of topic vocabulary to generate more expressive topic words. NTM is a typical representative of this kind of method [9]. NTM is based on the variational automatic encoder (VAE), which uses a continuous potential variable Z as an intermediate representation to induce potential topics in neural networks and make them more semantic continuous. On the basis of LDA topic model, Akash Srivastava [10] proposed a new neural topic model named prodlda in 2017, which can effectively improve the topic semantic continuity and better combine with neural network. Although the neural topic model can extract potential topic features more effectively, it loses the above information of text granularity. Therefore, in order to make up for the deficiency of neural topic model in text classification, this paper proposes a text expression model combining convolutional neural network and ProdLDA model based on neural topic model, and uses convolutional neural networks to extract text features, which effectively expands the text features, and achieves good classification results on long text data sets and short text data sets.

The work of this paper mainly includes the following aspects:

1) The text data is modeled based on neural topic model, and the text features are extracted. After modeling by word vector, the text features are extracted by convolutional neural network, and the two feature words are connected through the connection layer. Then the classification layer is used to classify.

2) After modeling the neural topic model, a new topic introduction method is proposed to make the topic features more semantic continuity and validity, and better connect with the feature matrix after convolutional neural pooling layer.

3) The algorithm is verified on two datasets. Compared with different classification algorithms, this ProdLDA-CNN model can effectively improve the text classification effect on short text data sets.

2. The text classification algorithm based on topic model and convolutional neural network

In order to improve the integrity and accuracy of text classification, this paper combines convolutional neural network and ProdLDA. After CNN mining the fine expression of text from word granularity, ProdLDA model represents the whole semantics of text through theme distribution from text granularity, so as to improve the classification effect of convolutional neural network.

2.1. Pre-training of the word vector

Before constructing ProdLDA topic model, we need to train the word vector. The word vector algorithm is to train words into vector form, which is easy to understand and calculate by computer. At present, there are mainly two kinds of word vector algorithms, namely one-hot coding and word2vec word vector algorithm.

One-hot coding, according to the position of words in the probability distribution, the components of vectors are represented by 1 or 0. For example, when the words in the document are too long, the dimension of the word vector is too high and sparse, which is easy to cause dimensional disaster, and can not maintain the context between words. Word2vec algorithm was proposed by Google in 2013, which overcomes the shortcomings of one-hot coding. The algorithm maps each word into k-dimensional dense and continuous real number vector by training, and judges whether there is semantic correlation between two words through the calculation of spatial distance. It can not only avoid dimensional disaster, but also extract semantic information in documents, and can significantly reduce the depth of convolutional neural network and its training costs.

There are two training models for word2vec word vector: Cbow and Skip-Gram. The former can predict focus words according to context words, while Skip-Gram model can predict context words according to focus words. In this paper, after the corresponding preprocessing of the text, the Cbow model in word2vec is used to train the text data, and the text data is transformed into the text vector matrix as the word vector representation of the input layer of convolutional neural network, aiming to do the next step of convolutional layer processing.

2.2. ProdLDA topic model

If only using word vector algorithm to train text, the whole semantic information of text may be lost. For example, two documents in the dataset do not have the same words, but they may be highly related semantically. Although the topic model can find the corresponding topic through the statistical analysis of the text, it ignores the context relevance between words. Therefore, we need to process the whole text semantics through convolutional neural network to obtain better classification effect. Prodlda topic model improves the traditional LDA model, can effectively improve the semantic continuity of topics, and can be more effectively combined with convolutional neural network [10]. Due to large-scale text training, the hidden topic and semantic information in the text can be obtained.

ProdLDA model introduces variational auto encoder , which is a generation model based on neural variational reasoning framework. Figure 1 shows the structure of the neural topic model. The parameter d represents a document to be processed, which is the probability distribution of the topic of the document, which is a k-dimensional vector. The topic generation process of the model is as follows:

1) Hidden variables $z \sim N(\mu_0, \sigma_0^2)$ are selected, where n is Gaussian distribution and Z is independent random variable.

2) The topic distribution of the text is set to:

$$\theta_d = soft\max(W_\theta z) \tag{1}$$

3) For the nth word in the document, there is a derivation formula (2). And W_ϕ and W_θ are trainable variables.

$$w_n = soft\max(W_\phi \theta_d) \tag{2}$$

In this case, the likelihood function of document d is as follows:

$$p(d \mid \mu_0, \sigma_0) = \int_z p(z|\mu_0, \sigma_0^{\,2}) \prod_{\mathrm{n}} p(w_n \mid \theta) p(\theta \mid z) dz \tag{3}$$

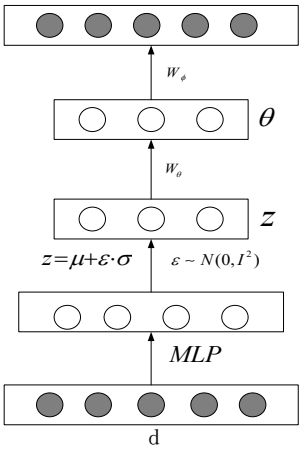


Figure.1 The ProdLDA topic model

ProdLDA model introduces variational inference network, which approximately represents the posterior probability. Among them, MLP is independent. This is the biggest difference between prodlda model and traditional LDA model. Therefore, ProdLDA model can better combine with convolutional neural network, and the vector dimension is the same when feature splicing, which is very convenient and efficient. The loss function of ProdLDA model is shown in equation (4).

$$L_u = -E_{q(z|d)}[\sum_{n=1}^{N}\log(p(w_n \mid \theta)p(\theta \mid z))] + D_{KL}[q(z \mid d)p(z \mid \mu_0, \sigma^2)] \tag{4}$$

The first item is the reconstruction loss. The reconstruction process is to generate the corresponding documents from the hidden variable Z. The second item ensures that the posterior probability $q(z \mid d)$ learned by the model is as close as possible to the actual posterior probability. Where, D_{KL} represents Kullback-Leibler divergence, which is used to evaluate the similarity of the distribution of two topics. In the experimental coding, we use the repeated parameter technique[13] $q(z \mid d)$ to sample z from it. For details, see formula (5), where the Gaussian distribution N (0, I) is sampled.

$$\hat{z} = \mu(d) + \hat{\in} \bullet \sigma(d) \tag{5}$$

The innovation of ProdLDA model is that the vector matrix of subject features is not constrained by polynomials before splicing, and the weighted product [10] of experts is used to replace the polynomial of this word vector splicing, which can be seen from formula (6).

$$p(W \mid \theta, \beta) \propto \prod_k p(w_n \mid z_n = k, \beta)^{\theta_k} \tag{6}$$

The ProdLDA topic model is a neural network in essence, and its feature extraction ability is stronger. Therefore, in this paper, the CNN neural network is used to extract the overall features of the text. After the pooling layer, the topic features extracted from ProdLDA model are spliced with them, and then input into the classification layer together.

2.3. The convolutional neural network

Convolutional neural network (CNN), proposed by Krizhevsky in 2012, has made remarkable achievements in natural language processing, image processing and other fields. The structure of convolutional neural network mainly includes convolutional layer, pooling layer (sampling layer) and output layer. The connection between neurons is incomplete, which can fully learn local features and make feature extraction more detailed. In the convolutional layer, the connection parameters between the neurons of convolutional kernel are shared, which reduces the complexity of the neural network model and greatly reduces the training cost; while the pooling layer is mainly responsible for extracting significant features and reducing the dimension of data.

It is precisely because the neurons in convolutional network share weights, which is very conducive to the acceleration of GPU parallelization, training speed is very fast, good at processing long text tasks, and can quickly extract spatial structure features. Therefore, this paper uses convolutional neural network to extract the overall features of the text, and after the pooling layer, the topic features extracted by ProdLDA topic model are spliced. After the stitched feature vectors are input and output layer (Softmax), the fused convolutional neural network text classification model is generated after parameter optimization and weight adjustment. In the experiment, we use the relu function as the activation function of the convolutional neural network.

2.4. The flow of the text classification algorithm based on ProdLDA model and CNN

In this paper, we introduce a neural topic model ProdLDA into convolutional neural network, and construct the CNN text classification algorithm based on ProdLDA topic model. Firstly, the potential topic features of the text are extracted by ProdLDA topic model, and the model is generated to construct the topic probability distribution of the document; secondly, the convolutional neural network is used to extract the text features of the word granularity; finally, the topic features are spliced into the semantic features extracted by CNN through the probability distribution of the topic, and the assembled feature word matrix can be used as the input of the full connection layer and merged into a table. After reaching the model, the classification can be carried out. Finally, the text classification algorithm based on ProdLDA topic model and convolutional neural network is obtained (ProdLDA-CNN).

The flow of Prod-LDA CNN model is as follows:

1) After processing the text data set, the document data set is transformed into word vector by using word2vec algorithm.

2) Initialize the key parameters of ProdLDA topic model, such as selecting the optimal number of topics, etc.

3) The ProdLDA model is trained iteratively. In the training process, the point product operation is added to calculate the text similarity between word vectors, and the feature information of word granularity is better extracted. The feature matrix B (i.e. the word granularity feature information extracted by ProdLDA model) is constructed to evaluate the performance of the model.

4) The probability distribution of documents is obtained by ProdLDA topic model.

5) The convolution neural network is trained to extract the whole feature information from the text granularity and construct the feature matrix A.

6) Merge the representation model. After the pooling layer of the convolutional neural network, the text features extracted from the convolutional neural network and the word granularity features extracted from the ProdLDA topic model are spliced to obtain a new feature vector, that is, the feature matrix A and the feature matrix B.

7) After training convolutional neural network, the feature vector of the spliced text is input into the output layer (classification layer). After parameter optimization and weight adjustment, the ProdLDA-CNN model is established by fusing ProdLDA topic model.

3. Experimental results and analysis

3.1. Experimental datasets

In the short text classification experiment, this paper uses 15 categories, 380000 pieces of data and 9 categories of 200000 pieces of data in today's headlines Chinese news data set do a comparative experiment (https://github.com/skdjfla/toutiao-text-classfication-dataset). Today's headlines data set is a short text data set collected from today's headlines client in 18 years. The data scale is 382688 pieces, which are distributed in 15 categories, covering various popular neighborhoods. Some of the classifications are closely related, which increases the requirements for classification accuracy, such as story, culture, travel and other categories with more cross information.

3.2. Evaluation index

Commonly used in text classification, the commonly used evaluation indicators are: accuracy, precision, recall, F1 value and so on. In order to verify the effectiveness of the algorithm in text classification, we calculate the four indexes respectively. These evaluation indexes can be used to evaluate the classification effect in different dimensions.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\Pr = \frac{TP}{TP + FP} \tag{8}$$

$$\mathrm{Re} = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{10}$$

Among them, TP: true positive means that the prediction is correct. TN: true negative, indicating the negative example of correct prediction. FP: false positive, indicating that the prediction is positive, but the actual case is negative. FN: false negative, which means that the prediction is negative and the actual case is positive.

3.3. Experimental comparison and parameter setting

When the ProdLDA topic model is used to model news data sets, the number of topics will affect the feature extraction of the topic model, and the fitting performance will be greatly affected in the training process. Therefore, it is necessary to design the number of topics to make the theme features of the model best. In this paper, ACC is used as the evaluation standard of the model, and the gradient of 20 is used to design a comparative experiment under Sogou data set. The number of optimal topics is selected as 70 from the following Fig.2.

In convolutional neural network, in order to extract text feature information, 512 convolution kernels with size of 3 are set in convolution layer. In order to prevent over fitting, regularization is used to optimize the parameters of neural network.

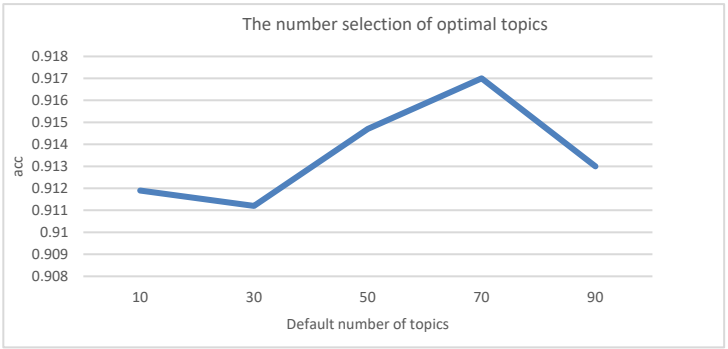


Figure.2 The number selection of optimal topics

In order to verify the effectiveness of the text model, we compare the classification model word2vec using convolutional neural network . In order to verify the validity of the model, the convolution layer adopts the same processing method as the model in this paper, and uses the same word vector training to train the same word vector as the convolution layer, which is used as the convolution layer after two full connection layer training Input, the experiment is carried out under the condition that other parameters are consistent.

And TMN (Ji Chuan Zeng, 18), which is based on neural topic model and memory network, also uses neural topic model (NTM) to optimize the classification model, and improves the model by combining the topic memory mechanism to obtain better classification effect. Compared with the model, it can verify the effectiveness of this model in text classification in different dimensions.

3.4. Experimental results

Table1 and Fig.3 are mainly about the short text classification experiments under different categories under the Chinese news data set of today's headlines. It can be seen from table 1 that three different models perform well in the classification task with the number of categories being 8, and all of the four evaluation indexes have achieved high accuracy. The accuracy of the first mock exam for CNN+word2vec is 90.4%, and the accuracy of the model is more than 1.3% and 1% respectively. The accuracy of the TMN model is improved. It is proved that the traditional convolutional neural network can achieve good results when the number of categories is small and the interference degree of noise information is low. The introduction of ProdLDA can effectively supplement the potential topic features and improve the classification accuracy.

Table.1 Short text classification experiment under 8 categories

	Acc(%)	Precision(%)	Recall(%)	F1(%)
Word2vec-CNN	0.9040	0.9058	0.9040	0.9041
TMN	0.9142	0.9149	0.9142	0.9143
ProdLDA-CNN	0.9175	0.9178	0.9131	0.9150

Fig.3 is the first mock exam of 15 headlines in Chinese headlines. From this, we can get that when the number of datasets increases, the noise information of corpus increases, and the information between categories is more. The average accuracy of the optimized single model CNN-word2vec is reduced. In the first mock exam, the ProdLDA-CNN model is still more accurate than the TMN model. The proposed model ProdLDA-CNN is improved by about 3% compared with the CNN-word2vec model without the neural theme model. It is proved that the accuracy of the single model can be improved by introducing the topic feature, and the rich feature of the topic information can effectively distinguish the more categories of the cross information under the complex corpus. The effectiveness of the proposed ProdLDA-CNN is proved.

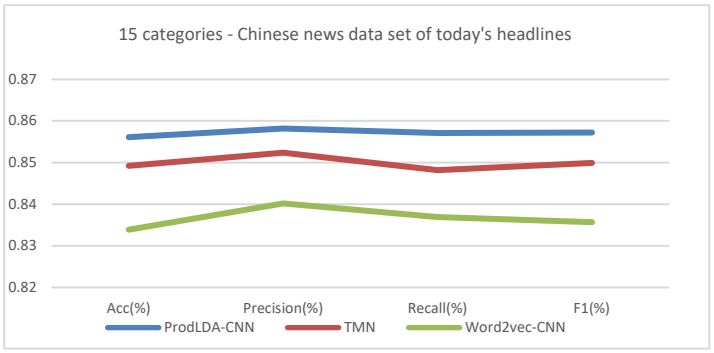


Figure.3 Short text classification experiment under 15 categories

4. Conclusion

Based on the neural topic model ProdLDA and convolutional neural network, this paper proposes ProdLDA-CNN model to do text classification. The convolutional neural network is used to extract the granularity features of high-dimensional text, and the neural topic model ProdLDA is used to extract the potential topic features. Then, the connection layer is established to connect the two text features. Finally, the classification level is processed.

Compared with different classification algorithms, this ProdLDA-CNN model can effectively improve the text classification effect on short text data sets. It is proved that the neural topic model can effectively enrich the topic feature information and make the classification result more accurate.

References

[1] Salton, G., A. Wong, and C.-S. Yang, A vector space model for automatic indexing.Communications of the ACM, 1975. 18(11): p. 613-620.
[2] Wiemer-Hastings, P., K. Wiemer-Hastings, and A. Graesser. Latent semantic analysis. in Proceedings of the 16th international joint conference on Artificial intelligence.2004.Citeseer.
[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learnin- g research, 2003, 3(Jan): 993-1022
[4] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, et.al. A Biterm Topic Model for Short Texts [J]. WWW 2013, 2013, Rio de Janeiro, Brazil.ACM 1445-1455
[5] Hofmann, T. Probabilistic latent semantic analysis. in Proceedings of the Fifteenth confe- rence on Uncertainty in artificial intelligence. 1999. Morgan .Kaufmann Publishers Inc.
[6] KIM Y． Convolutional neural networks for sentence classification［J / OL］． Computer Science，2014.[2019-05-01].https: / / arxiv.org / abs /1408.5882．
[7] Cho K, Van M B,Gulcehre C, et al. Learning phrase representations using RNN Encoder Decoder for statistical machine translation[J]. EprintArxiv, 2014.
[8] ZHOU C，SUN C L，LIU Z Y，et al.A C-LSTM neural network for text classification[J / OL].Computer Science,2015.［2019-05-01］.http: / / arxiv.org / abs /1511.08630．
[9] Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural Variational Inference for Text Proc -essing. In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY,USA, pages 1727–1736.
[10] Akash Srivastava. 2017. Autoencoding Variational Inference For Topic Models. arXiv : 1703. 01488.
[11] Jichuan Zeng , Cuiyun Gao , Irwin King.2018 . Topic Memory Networks for Short Text Classification.[cs.CL]. arXiv:1809.03664.