Research Article

Text Data Mining Algorithm Combining CNN and DBM Models

Rong Dai

Civil Aviation Flight University of China, School of Computer Science, Sichuan, China

Correspondence should be addressed to Rong Dai; dairong@cafuc.edu.cn

Received 18 August 2021; Revised 22 September 2021; Accepted 28 September 2021; Published 20 November 2021

Academic Editor: Sang-Bing Tsai

Copyright © 2021 Rong Dai. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The special text has a lot of features, such as professional words, abbreviations, large datasets, different themes, and uneven distribution of labels. While the existing text data mining classification methods use simple machine learning models, it has a bad performance on text classification. To solve this drawback, a text data mining algorithm based on convolutional neural network (CNN) model and deep Boltzmann machines (DBM) model is proposed in this paper. This method combines the CNN and DBM models with good feature extraction to realize the double feature extraction. It can realize the tag tree by constructing the tag tree and design the effective hierarchical network to achieve classification. At the same time, the model can suppress the input noise on the classification. Experimental results show that the improved algorithm achieves good classification results in special text data mining.

1. Introduction

With the in-depth study of text classification model in industry and academia, text information presents the phenomenon of exponential growth. Data mining of massive information to extract effective information has become a hot issue.

Text data mining is defined as the induction process of a single or multiple categories of objects based on different characteristics of the document data. Initially, text classification method is mainly used in a large number of naive Bayes machine learning methods [1]. Then, a series of machine learning algorithms including K-nearest neighbor algorithm [2], support vector machine (SVM), neural network [3], least squares [4], and decision tree [5] have been widely used in the field of text classification. Recently, the application of SVM has become a hot research direction for researchers in the field of text classification [6]. However, K-nearest neighbor algorithm, least square, and decision tree are used in simpler models with higher efficiency and can be optimized and improved based on these methods. Literature [7] proposed graph neural networks (GNN) based text inductive classification method. It first adopted GNN to learn fine-grained word representations based on their local structures and then aggregated word nodes into document

embedments to obtain classification results. Literature [8] proposed a new term weighting strategy, which makes more effective use of the nonoccurrence information of terms. The proposed weighting strategy also performs intraclass document scaling to better represent the discrimination ability of terms appearing in different document numbers of the same number of classes.

In terms of improving generalization ability, a selective integration theory is proposed, which has achieved good results for text classification applications. However, these models are the kind of shallow methods. When the data to be processed are massive and high-dimensional data, its classification is called complex data classification. When faced with complex data classification problems, the limitations of the algorithm based on this theory will be obvious. Specifically, the generalization ability is insufficient, and the requirements for text classification cannot be met. Therefore, how to obtain a deep machine learning method with strong generalization ability has become the mainstream of research.

The task of text classification can be divided into three steps: text preprocessing, text representation, and classification model construction. To manage the text information well, it is necessary to extract and classify text content scientifically and reasonably. The previous text

representation is generally in the form of a count. This approach has two drawbacks. First, this method needs to assume that words are independent from each other, but the actual words are all related to each other, which leads to the neglect of the text language between words. Second, the interference of human factors is required when selecting features, which results in the extracted features having quotient dimensions and sparseness, and the representation and generalization ability of the text are very poor. In addition, for special texts, it has a large number of professional words, abbreviations, large datasets, diverse topics, and uneven label distribution. A simple machine learning model is used by the existing special text categorization methods to reduce the performance of text categorization.

The contributions of this paper are as follows:

- (1) To overcome the shortcomings of the above-mentioned classification methods, a text data mining algorithm based on convolutional neural network (CNN) model and deep Boltzmann machines (DBM) fusion is proposed in this paper.
- (2) This method combines two models to achieve dual feature extraction. Tag tree is realized by constructing tag tree and designing effective hierarchical network.
- (3) At the same time, the effect of input noise on classification can be suppressed. It can effectively classify a large number of specialized words, abbreviations, and short text in documents and perform well.

The structure of this article is as follows. Section 2 introduces the general methods of big data mining. Section 3 introduces the recommendation algorithm. Section 4 shows the experimental results and analysis.

2. Common Text Classification Methods in Data Mining

There are two main methods of text classification: rule-based and statistics-based. At present, popular machine learning methods mainly include support vector machine (SVM), logistic regression, naive Bayes, decision tree, K-nearest neighbor, artificial neural network, integrated learning, tag correlation classification, and hierarchical classification methods.

Support vector machine (SVM): the principle of SVM is to find a hyperplane to meet the classification requirements so that the points in the training set can be separated from the classification plane as far as possible. When it comes to convergence training on big data, the SVM speed is very slow. It needs large concurrent computers and equipment resources with large storage capacity to support it. However, its advantage lies in that it can well overcome the influence of sample distribution so that the experimental effect is very good and the generalization ability is very good. SVM is a shallow linear model to classify different data. If the low-dimensional data vector space cannot be classified, the mapping method can be used to find the best hyperplane.

Logistic regression (LR): the LR model selects parameters according to the input variable Z and calculates the output variable, that is, $h_{\theta}(z) = p(y = 1/z; \theta)$ is the possibility of 1. Here, the assumption of logistic regression model is shown in equation (1), and sigmoid function of type S is shown as follows:

$$h_{\theta}(z) = \frac{1}{1 + e^{-\theta^T z}}, \quad (1)$$

$$g(x) = \frac{1}{1 + e^{-x}}. \quad (2)$$

The logistic regression model is $h_{\theta}(z) = g(\theta^T z)$ obtained from equations (1) and (2), where z is the eigenvector of the classification target.

3. Improved Text Data Mining Method Based on CNN and DBM

3.1. The Text Classification of CNN and DBM. CNN is a deep learning model characterized by weight sharing, which is an extension of BP neural network [9, 10]. CNN uses the gradient descent method to achieve weight adjustment. Its characteristic is that the direction of weight adjustment is toward the direction of the fastest gradient, which improves the network convergence speed. In terms of feature mapping, if the weights of neurons are consistent, the parallel learning of the network can be realized, which is a feature that CNN differs from neural network.

DBM is the basic modeling unit of network, and it is a model structure composed of RBM with undirected graph connection. The schematic diagram of DBM can be seen in Figure 1. It is mainly composed of unsupervised pretraining and supervised fine-tuning [11], which are roughly consistent with network results when selecting network nodes. DBM is able to efficiently combine local and global feature information [12]. It consists of a set of multiple visual units that make up its input layer (v). The hidden layer (h) consists of a number of hidden cells in sequence and finally the output layer, which constitutes the DBM model. Adjacent layers in the model are connected by undirected graphs.

There are three main advantages of DBM. First, the weight can be updated through prior knowledge, which can extract features well. Second, the weight is updated by prior knowledge, which can effectively suppress input noise. The third is to simultaneously sample and calculate the weights of the neighboring nodes [13, 14]. This will give a more accurate text representation. DBM also has its own shortcomings, mainly in the expansion of the number of network layers, and the number of connected nodes increases; the computational complexity is exponential.

3.2. The Improved Text Classification Method. An improved text categorization method proposed in this paper is based on the improvements of CNN and DBM models. There are three steps to improve the CNN and DBM model components. The overall framework is shown in Figure 2.

FIGURE 1: The schematic diagram of DBM.

FIGURE 2: Framework of CNN and DBM models.

In order to improve the accuracy of CNN and DBM model classification, the third step of this model, namely, hierarchical classification, is improved.

Figure 3 is a detailed framework diagram of the improved model, and the middle part is the feature extraction layer. In this step, CNN model is adopted to realize local feature extraction y_c and global feature y_e supplement for the input text representation. Then, DBM is used to fuse the two features and finally classify them.

In this framework, the output of CNN is the extracted local text feature y_c , and y_e is the entity feature which is the global feature. The input dimensions of both are the same, and they constitute the input formula of the DBM model, which is expressed as follows:

$$V = \frac{1}{1 + e^{Cy_c + Ey_e}}. \quad (3)$$

Then, each time a layer of hidden layer is passed, the corresponding weight w_i is obtained. After model training, pretraining and fine-tuning, model testing, and finally the label of the target sample. In addition, in order to speed up the training speed of the model, ReLu activation function is adopted for training in this paper.

3.3. Hierarchical Classification. The output tag classification of DBM model is realized by designing label tree hierarchy (LTA). The LTA classifies the labels in a tree structure and renames the labels in a sequence of tree structures to form new labels. According to the characteristics of the experimental dataset adopted in this paper, all the labels are divided into two layers for layer processing. The first layer is a rough classification, which corresponds to the father node. The second layer is a fine classification, which corresponds to a leaf node.

There will be some errors through this layering. The error is obtained by the difference between the model classification and the real classification. At the same time, it is fed back to the CNN network in the first step of the model. The model receives the feedback error and corrects it and adjusts the weight until accurate classification is achieved.

3.4. The Evaluation Index of Text Classification Performance. The evaluation index of text classification method is based on the prediction of text classification. In general, there are three categories of indicators, namely, basic indicators, macro- and microaverage indicators, and ROC curve indicators.

The basic performance indicators of text classification include accuracy rate P , recall rate R , measure value F , and similarity S . The accuracy rate P is a measure of the precision of the retrieval system and is defined as follows:

$$P = \frac{\text{related files retrieved by the system}}{\text{total number of files retrieved by the system}}. \quad (4)$$

The recall rate R is a measure of the entire document system and is defined as follows:

$$R = \frac{\text{related files retrieved by the system}}{\text{total number of related files in the system}}. \quad (5)$$

Select the measure value $F1$ as the classification index and it is the weighted harmonic average of equations (2) and (3).

$$F_1 = \frac{2PR}{P + R}. \quad (6)$$

The similarity S is defined as

FIGURE 3: Framework for improved model.

$$S = \frac{\text{related files retrieved by the system}}{\text{system retrieved files} + \text{related files not retrieved}} \quad (7)$$

The basic indicators P and R of classification performance are indicators that measure a certain category. The metrics for the entire dataset are macro and micro. The macroaverage reflects the overall performance of the algorithm, while the microaverage reflects the overall arithmetic performance of the algorithm. These performance indicators are expressed as follows”

$$\begin{aligned} \text{Macro} - P &= \frac{\sum_{K=1}^K P_K}{K}, \\ \text{Macro} - R &= \frac{\sum_{K=1}^K R_K}{K}, \\ \text{Macro} - F_1 &= \frac{\sum_{K=1}^K F_{1K}}{K}, \\ \text{Micro} - R &= \frac{\sum_{K=1}^K A_K}{\sum_{K=1}^K A_K + \sum_{K=1}^K C_K}, \\ \text{Micro} - F_1 &= \frac{2 * \text{Micro}P * \text{Micro}R}{\text{Micro}P + \text{Micro}R}, \end{aligned} \quad (8)$$

where K represents the number of categories divided. It can be seen that the macroaverage has the characteristics of weight sharing, and the weight of each category is the same. The arithmetic mean of microaverage makes it more susceptible to large categories.

The ROC curve is a comprehensive indicator of continuous variables of sensitivity and specificity. If the area formed by the indicator curve is larger, it reflects the higher accuracy of the algorithm.

4. Experimental Results and Analysis

Dataset 1 used in this experiment is a performance comparison document on the medical dataset. The total number of samples is 9666, which is divided into 39 categories, and the corresponding types are multiclass. Dataset 2 is selected from the dataset. The total number of samples is 1000, which is divided into 168 categories. The corresponding types are multicategories. Dataset 3 is selected from the dataset. The total number of texts reaches 1000000, which is divided into 150 categories. The corresponding type is multilabel. These three kinds of experimental objects can better extend and verify the generalization ability of the model proposed in this section. Table 1 is a comparison of several datasets.

In the experimental subjects selected in this paper, the ratio of the training samples to the test samples of the subjects is 7:3. In addition, the sliding window step size of the CNN model is set to 50 to shift, so as to avoid changing the meaning of the representative word.

The performance indicators of different features of dataset 1 are compared. The specific data indicators are shown in Tables 2 and 3. From the data of different indicators, it can be concluded that the performance of the improved model proposed in this paper is better than that of other models regardless of BOW+ or DSE features. As far as the shallow model approach is concerned, the BOW+ has better performance than the DSE, while the DSE represents better performance for the model proposed in this paper and the improved CNN model.

Through experiments on different models and feature representation methods of dataset 2, Tables 4 and 5 show that the performance of the improved model proposed in this paper is superior to other models under different feature representation methods. For the shallow model method, BOW+ has better performance than DSE, while DSE represents better performance for the model proposed in this paper and the improved CNN model.

TABLE 1: Dataset information.

Dataset	Number of samples	Number of categories	Type
Dataset 1	9666	39	Multiclass
Dataset 2	1000	168	Multi categories
Dataset 3	1000000	150	Multilabel

TABLE 2: Performance comparison of BOW+ features in dataset 1 (%).

	P	R	F	S
PM	21.71	31.55	25.88	40.56
LDA	55.22	54.67	54.85	52.16
SVM	57.44	57.06	57.25	53.93
NB	53.22	55.01	54.02	52.09
LR	54.06	53.22	53.99	54.09
MTI	62.07	63.01	62.07	57.01
MU	63	59	61.99	56.09
CNN_{flat}	53.11	53.66	52.99	52.22
CNN_H	60.01	58.99	59.22	55.99
$C - B_{flat}$	54	52.33	53.99	51.89
Proposed	64.97	63.2	64.12	59.22

TABLE 3: Performance comparison of DSE features in dataset 1 (%).

	P	R	F	S
PM	21.77	31.55	25.88	40.56
LDA	50.99	50.98	50.99	50.56
SVM	47.55	47.66	47.88	48.09
NB	50.01	50.99	50.06	51.01
LR	50.01	46.09	46.99	48.01
MTI	52.01	52.06	51.77	57.99
MU	50.07	50.09	50.66	50.99
CNN_{flat}	55.09	52	53.09	54.01
CNN_H	63.99	61.99	62.36	60
$C - B_{flat}$	58.33	57.33	58.99	54.36
Proposed	68.99	66.33	67.66	61

TABLE 4: Performance comparison of BOW+ features in dataset 2 (%).

	P	R	F	S
PM	25.17	36.51	30.17	42.07
LDA	47	46.33	47	48.36
SVM	48.14	49.22	49.22	49.33
NB	46.33	46.22	46.36	48.39
LR	45.2	45.69	45.33	48.33
MTI	52.33	51.74	51.22	51.01
MU	51.77	52.88	52.11	51.85
CNN_{flat}	46.33	42.04	44.2	47.2
CNN_H	61.22	59	60.33	56.88
$C - B_{flat}$	55.22	52.99	53.55	52.99
Proposed	64.99	62.89	63.87	60.85

For the analysis results in Tables 2–5, more datasets can be obtained for the sample data. BOW+ and DSE feature representation is used to conduct performance comparison

TABLE 5: Performance comparison of DSE features in dataset 2 (%).

	P	R	F	S
PM	25.66	36.11	30.17	42.08
LDA	40.22	39.22	39.33	45.02
SVM	40.12	40.88	40.22	45.99
NB	40.12	40.89	40.87	45.96
LR	38.33	35.22	36.74	40.22
MTI	42.1	42.09	42.77	46.22
MU	39.44	39.25	39.88	55.44
CNN_{flat}	48.77	48.55	48.52	49.52
CNN_H	66.99	63.44	65.28	59.66
$C - B_{flat}$	59.77	58.24	58.96	57.21
Proposed	71.22	67.22	69.01	62.01

TABLE 6: Performance comparison of BOW+ features in dataset 3 (%).

	P	R	F	S
LDA	25.49	36.83	30.4	42.44
SVM	47.32	46.65	47.23	48.73
NB	48.46	49.54	49.45	49.7
LR	46.65	46.54	46.59	48.76
MTI	45.52	46.01	45.56	48.7
MU	52.65	52.06	51.45	51.38
CNN_H	52.09	53.2	52.34	52.22
$C - B_{flat}$	46.65	42.36	44.43	47.57
Proposed	61.54	59.32	60.56	57.25

TABLE 7: Performance comparison of DSE features in dataset 3 (%).

	P	R	F	S
LDA	51.2	47.7	49.63	49.87
SVM	48.1	48.3	48.99	49.99
NB	45.3	43.2	44.12	47.21
LR	52.1	49.2	50.12	50.78
MTI	55.2	47.2	51.85	50.21
MU	53.9	49.1	51.96	50.21
CNN_H	56.3	55.1	55.85	53.21
$C - B_{flat}$	49.3	46.1	47.14	49.12
Proposed	68.1	66.1	67.14	61.77

It can be concluded from the index data that the improved model proposed in this paper has better performance than other models.

experiments for the 9 models, and the experimental results are shown in Tables 6 and 7.

ROC performance experiments were conducted on three datasets. ROC curves of five models on three medical abstract datasets are shown in Figures 4–6. The abscissa is specificity and the ordinate is sensitivity. The closer the curve is to the upper left, the better the performance. It is not difficult to see from the figure that the improved method has the best performance on medical, BioTex and Medline datasets.

Figures 4–6 show ROC performance comparison of different models on SVM, LDA, CNN_H , $C-B_{FLAT}$, and the improved method dataset. We can draw a conclusion from the pictures. Firstly, deep learning models are better than shallow learning methods. Secondly, hierarchical classification is better than flat classification. Thirdly, the improved model proposed in this paper can have optimal performance for different datasets.

FIGURE 4: ROC curves on dataset 1.

FIGURE 5: ROC curves on dataset 2.

FIGURE 6: ROC curves on dataset 3.

5. Conclusion

According to the characteristics of special text, the existing special data mining method of text classification using simple machine learning model has poor performance. In order to solve this problem, a new improved data mining method based on CNN model and DBM model is proposed. This method combines CNN and DBM models with good feature extraction to achieve dual feature extraction. It can realize the reclassification of tags by constructing the tag level of tree structure and designing an effective hierarchical network. The model can suppress the influence of input noise on classification. The experimental results show that the improved model has good effect on the special domain text. The classification is a part of mining, and further information mining will be analyzed in the next research.

Data Availability

The labeled dataset used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported in part by the Civil Aviation Safety Capability Fund (0242008) and in part by the Key Laboratory of Civil Aviation Flight Technology and Flight Safety (FZ2020ZZ02) and National Natural Science Foundation of China Civil Aviation Joint Fund (U2033213) and National Key Laboratory Project "Research on Nonlinear Dynamic Characteristics of Helicopter High Power Density Gear Transmission System" (No. HTL-0-19K01).

References

- [1] K. V. Ranjitha and B. S. V. Prasad, *Optimization Scheme for Text Classification Using Machine Learning Naïve Bayes Classifier*, pp. 576–586, Springer, Singapore, 2020.
- [2] Y. Meng and W. Gang, "File type detection algorithm based on principal component analysis and K nearest neighbors," *Journal of Computer Applications*, vol. 36, no. 11, pp. 3161–3164, 2016.
- [3] J. Xu, Y. Cai, X. Wu et al., "Incorporating context-relevant concepts into convolutional neural networks for short text classification," *Neurocomputing*, vol. 386, pp. 42–53, 2020.
- [4] Z. Yi, Z. Zhen, and Z. Qing, "Parallel nonlinear least squares for high dimensional multi-class data classification," *Journal of Chinese Computer Systems*, vol. 35, no. 3, pp. 579–583, 2014.
- [5] W. Yang, Z. Ming, and J. Yan, "Study on short text categorization Technology oriented towards field of auditing," *Microelectronics & Computer*, vol. 1, pp. 5–10, 2015.
- [6] H. Lv and L. Fang, "Application of text classification for specific domains based on combination kernel function SVM," *Computer Systems & Applications*, vol. 5, no. 5, pp. 124–128, 2016.
- [7] Y. Zhang, X. Yu, and Z. Cui, "Every document owns its structure: inductive text classification via graph neural networks," 2020, <http://arxiv.org/abs/2004.13826>.
- [8] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," *Journal of Informetrics*, vol. 14, no. 4, p. 101076, 2020.
- [9] G. Dong, L. Xiao, and Z. Qiu, "Internet short-text classification method based on CNNs," *Computer and Modernization*, vol. 4, pp. 78–81, 2017.
- [10] L. Xing, W. Jiang, and X. Ke, "Novel image feature extraction algorithm based on fusion Auto Encoder and CNN," *Application Research of Computers*, vol. 34, no. 12, pp. 3839–3842+3847, 2017.
- [11] M. Zhou, Y. Wang, Z. Tian, Y. Lian, Y. Wang, and B. Wang, "Calibrated data simplification for energy-efficient location sensing in internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6125–6133, 2019.
- [12] Z. Jian, Z. Zhi, and C. Ke, "Review on development of deep learning," *Journal of Jiangsu University*, vol. 36, no. 2, pp. 191–200, 2015.
- [13] G. Li, S. Tie, and H. Xi, "An approach to the text categorization of the Kazakh language based on SVM-modified KNN algorithm," *Journal of Northwest Normal University*, vol. 3, pp. 48–53, 2014.
- [14] M. Zhou, Y. Li, M. J. Tahir, X. Geng, Y. Wang, and W. He, "Integrated statistical test of signal distributions and access point contributions for Wi-Fi indoor localization," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 5, pp. 5057–5070, 2021.