Journal of Computers Vol. 31 No. 1, 2020, pp. 253-260
doi:10.3966/199115992020023101023

Power Data Classification Method Based on Selective Ensemble Learning

Yi-Ying Zhang[1], Fei Liu[1*], Hao-Yuan Pang[1], Bo Zhang[2], Yang Wang[3]

[1] College of computer science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300457, China

[2] Global Energy Interconnection Research Institute co. Ltd., Nanjing 210003, China

[3] Tianjin Electric Power Company, State Grid Corporation of China, Tianjin 300010, China
1006124568@qq.com

Received 13 January 2019; Revised 17 March 2019; Accepted 17 March 2019

Abstract. The power data implies a large number of user's characteristic attributes and the user's power consumption rules. If these potential behavior attributes of the user can be mined in this way, the precise power supply on the power supply side will provide strong support. In this paper, based on the user's electricity information data, the improved TF-IDF is used to preprocess the data. The whole two-layer ensemble learning framework is adopted, and the word vector is introduced to expand the characteristics of the text. Finally, the result of the first layer is obtained. After the feature splicing with the word vector, the classification prediction is performed through the CNN network, and the final prediction model is obtained to predict and classify the user's power usage behavior. Compared with the traditional CNN model, the classification effect of this paper has been significantly improved.

Keywords: CNN, ensemble learning, TF-IDF, word vector

1 Introduction

With the continuous development of smart grid technology, the amount of data generated in the power grid has become larger and larger. If valuable information is extracted from these data, it becomes an urgent problem to be solved. It is also very important for the user to use the electricity data mining. On the power side, through the data analysis and mining of the users of the electricity, it is more accurate to formulate the corresponding power consumption strategy for the user, which can make the power grid unit achieve the purpose of precise marketing.

At present, there are many techniques and algorithms applied to text classification, such as naive Bayesian algorithm, K nearest neighbor algorithm, neural network, support vector machine, and the like. Among them, the SVM classification algorithm has good generalization ability and learning ability, but the traditional SVM classification algorithm is susceptible to data sets, classifiers and training parameters. In this paper, according to the influence of classifier and training parameters, according to the characteristics of power data, the corresponding classification strategy is formulated, and the power data is accurately classified [1].

Selective ensemble learning can handle the above problems very well. For the text classification of power data, this method combines the multi-layer model framework to more accurately mine the potential value information of users from the power data. Compared with the traditional classification algorithm, the classification accuracy of this method is more accurate, and the generalization ability of the model is stronger. According to the user's potential information characteristics, the user can be classified reasonably, which can realize the precise marketing of the power grid unit and the reasonable dispatch of electric energy.

[*] Corresponding Author

According to the idea of selective ensemble [2], the selected base classifiers are selectively ensemble, and the better classifiers are selected for prediction. The power data classification method based on selective ensemble learning is studied. There are three main innovations in this paper: (1) The TF-IDF method is improved, and the S-TFIWF based on the Boolean model is proposed for the deficiency of TF-IDF; (2) Selective ensemble is used on the first layer of ensemble learning model, and each base classifier is selectively selected for ensemble, which improves the accuracy of the model; (3) In the second layer, the word vector is spliced with the result of the first layer, and then the final prediction model is obtained through the CNN network.

2 Related Work

In the field of classification and prediction, after the ensemble ideas such as Zhou Zhihua [3] are used in the field of classification and prediction, people are constantly using ensemble learning ideas for classification prediction. The research shows that the classification prediction method based on ensemble learning is better than the single classifier prediction method, but it has the problems of too long learning time and too large storage space. The emergence of selective ensemble ideas has solved this problem better. The experimental results show that the classification prediction method based on selective ensemble is better than the general classification based on ensemble learning. However, in the process of selective ensemble learning, the threshold setting and parameter selection are difficult to set. These parameters are difficult to set. The setting has a very large impact on the results. In addition, there are few researches on selective ensemble, there is no uniform definition of the selection and ensemble strategies of each base classifier, and there is no mature theoretical combination, so there is still a great space for exploration in the selective ensemble method.

In order to improve the efficiency of text categorization [4], an improved KNN algorithm based on clustering is proposed. Before the algorithm starts, the improved χ^2 statistic method is used to extract the text features. Then the text set is clustered into several clusters according to the clustering method. Finally, the improved KNN method is used to classify the clusters. Experimental comparison and analysis results show that this method can better classify texts.

However, there is no text classification prediction method based on selective ensemble learning. Selective ensemble learning can maximize the efficiency of each base classifier to achieve more accurate classification results. In this paper, the S-TFIWF algorithm based on Boolean model is used to preprocess the text data, and the two-layer selective ensemble model of Stacking is used to classify the data more accurately, which achieves a more accurate classification effect.

3 Data Classification Method Based on Selective Ensemble Learning

This paper adopts a two-layer ensemble learning framework. As shown in Fig. 1, because the data set is the user’s power consumption information and is text data, the data is processed in the data classification stage using the Boolean model-based S-TFIWF method. Weight the data. Then this paper use Stacking two-layer ensemble learning model [5-7], using a variety of algorithms to train data in the first layer of training, such as: random gradient descent, linear kernel SVM, naive Bayes and logistic regression, etc. When the training results are ensemble, the base learner is selected using the ranking method [8]. The result of the ensemble of the first layer is spliced with the word2vec feature [9] as the second input, and finally the final classification prediction is performed by CNN.

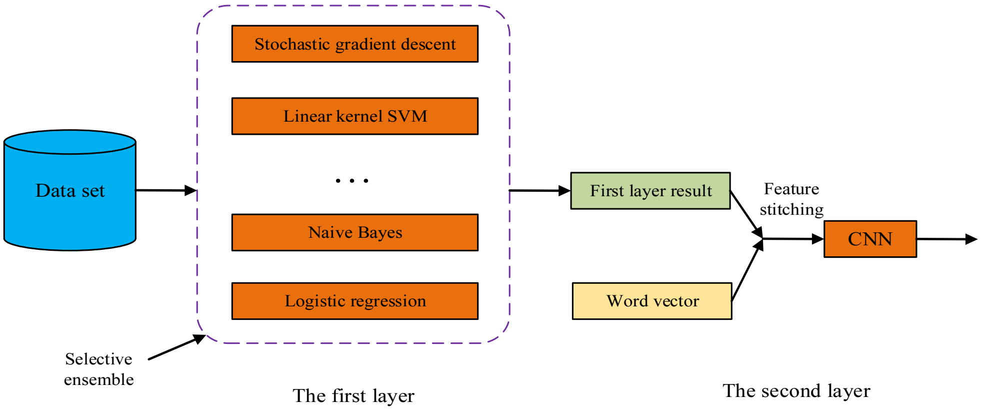

Fig. 1. Stacking ensemble learning framework

3.1 S-TFIWF Based on Boolean Model (Supervised TFIWF)

The TFIDF algorithm [10] is a classic text feature weighting method that measures the importance of a word in a document. TFIDF is proportional to the number of occurrences of a word in a document and inversely related to the number of occurrences of that word in the entire language. The formula of the TFIDF algorithm is shown in equation (1):

$$TFIDF(\omega,d) = TF(\omega,d) * \log(N / DF)(\omega)) \qquad \textbf{(1)}$$

In equation (1), TF(w, d) is the word frequency of the word w in the article d; N is the total number of documents in the corpus; and the number of documents containing the word w in the DF(w) training corpus.

Although the TFIDF algorithm comprehensively considers the influence of word frequency in the document and the distribution of words in the overall corpus, the TFIDF algorithm has many shortcomings, for example: the number of occurrences of words in the document TF has too much influence on the overall weight, and does not take into account the distribution of words between different classes and so on. Insufficient for TFIDF in this paper, we consider the influence of TF on the overall weight and the distribution of words between different classes. The S-TFIWF algorithm based on Boolean model is proposed. The specific formula is shown in formula (2):

$$S - TFIWF(\omega,d) = TF(\omega,d) * \log(\frac{N}{WF(\omega)}) * \sqrt{\sqrt{\sum_j (p_{wj} - p_w)^2} / \sum_j p_{wj}} \qquad \textbf{(2)}$$

In the formula (2), TF(w, d) is the occurrence or absence of the word w in the article d. Under the Boolean model, when the word w appears in the document d, TF(w, d) is 1, and vice versa; N is the sum of the occurrences of all words in the training corpus; WF(w) is the number of occurrences of the word w in the training corpus; the probability that the pwj word w appears in the class j; p_w is the average probability that the word w appears in all classes; log(N/ TF(w, d)) means that when a word appears multiple times in the overall corpus, the word may be a common word, which is small for the judgment of the category, and should be suppressed; the content with the root number indicates that the word is utilized. The probability variance between different classes represents the information of the distribution of the words w. Through this information, we can judge that when a word has a large variance in different classes, the word is more distinguishable and the weight should be bigger.

3.2 Stacking Selective Ensemble Strategy

Ensemble learning includes three main elements: the generation method, the base classifier, and the ensemble method. However, in the selective ensemble, a major element is added: the choice of the base classifier is not ensemble. Integrate all the base classifiers, but by formulating the corresponding selection strategy, only select the good base classifier for ensemble, and discard those poor base classifiers.

In this paper, the Stacking ensemble model is adopted. Stacking is divided into two layers of learning process, as shown in Fig. 1. Firstly, multiple algorithms are learned on the original data set, and then trained again on the data set composed of all the prediction results and the original sample real values. In this paper, we selectively integrate in the first layer of the Stacking model. The strategy is to use multiple classification methods in the first layer. In the first layer ensemble, the ranking method is used to select the base classifier. The ranking method is to classify all the bases. The device performs an evaluation sort and then selects the base classifier in that order. The biggest advantage of the ranking method is that the base classifier is fast.

When integrating the training results, the base learner is selected using the ranking method. The specificity of the ranking method is to use the Kappa coefficient to filter each base classifier. The screening rules are as shown in equation (3):

$$k_i = \frac{p_0 - p_i}{1 - p_i} \tag{3}$$

In equation (3), p_0 is the sum of the number of samples correctly classified by each base classifier divided by the total number of samples, that is, the overall classification accuracy; p_i is the number of samples correctly classified by the classifier divided by the total number of samples, usually the calculation of Kappa coefficient The result is -1~1. The calculation result of Kappa coefficient is between 0~1. The Kappa coefficients of the base classifier are sorted, and the base classifier with the precision above 0.8 is selected for ensemble.

3.3 Introduction of Word vector Word2vec and Feature Stitching

The word vector studies how to represent a word as a real number vector. word2vec is one of the training methods. It obtains the distributed expression of words based on the context information of the word. word2vec is an unsupervised method because the test set can be the use of annotation data to a certain extent alleviates the adverse effects of missing labels.

Suppose a document contains N words, each of which is represented by a K-dimensional word vector. Dimensional inconsistency caused by different document lengths cannot be resolved. Therefore, this paper adopts the following method to add and average the word vectors: If the expression ability of a certain word is stronger, the vector obtained by summing and averaging is biased toward the direction of the word vector, as shown in Fig. 2.

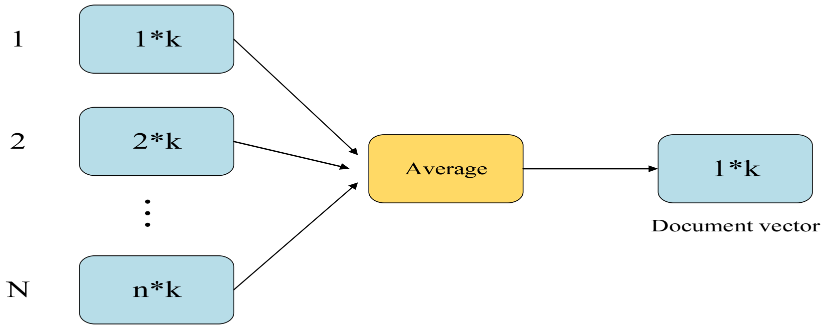

Fig. 2. Weighted average word vector

After obtaining the vector expression of the document under word2vec, it is not a good strategy to directly splicing the K-dimensional word2vec vector on the S-TFIWF feature, because the S-TFIWF feature has high sparsity and high dimension, and word2vec the dense form is too big. Combined with the Stacking ensemble framework introduced earlier, this paper spliced the word2vec feature in the second layer of Stacking. If the first layer has 20 classifiers and the word2vec dimension is K, then the dimension of the second layer feature becomes: K+20K+20. S-TFIWF+word2vec completes the expression of the document from the two perspectives of feature words and word vectors.

3.4 Convolutional Neural Network Classifier

After the first layer of ensemble training results and the training set word vector are spliced together, through CNN training, this paper constructs a four-layer convolutional neural network [11-13], as shown in Fig. 3.

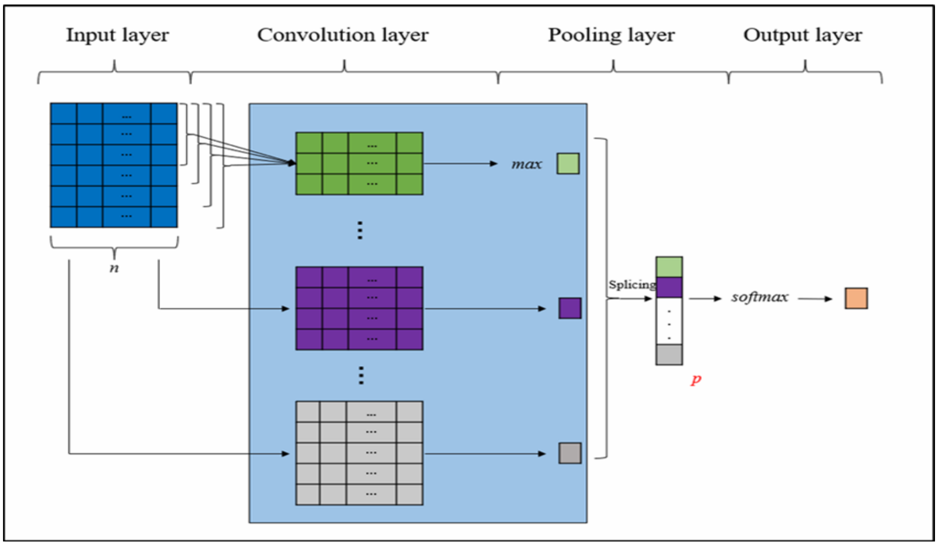

Fig. 3. Convolutional neural network used in this paper

The first layer is the input layer. The input layer is a matrix corresponding to the text to be classified, I $\in R^{s \times n}$, each row of the matrix represents a vector corresponding to each word in the sentence, the number of rows s is the number of words of the sentence, and the number of columns n is the dimension of the vector. In the training process, the stochastic gradient descent (SGD) method is used to fine-tune the word vector, the word vector of the number is fixed, and the other word vectors are fine-tuned to improve the generalization ability of the model.

The second layer is a convolution layer. A convolution matrix window W$\in R^{h\times n}$ having the same number of columns as I and a number of rows h is used, and each h row and n column matrix blocks of the input layer matrix I are sequentially subjected to convolution operations from top to bottom. Get the convolution result r_i, is:

$$r_i = W \cdot I_{i:i+h-1} \tag{4}$$

In the formula (4), i=1, 2, ..., s-h+1; $I_{i:i+h-1}$ represents the i-th h-row n-column matrix block from top to bottom; "··" Indicates the point multiplication operation, which is to multiply all the elements of the same position in the two matrices and then sum. In addition, there are multiple convolution windows in each group, and the element values between the respective convolution windows (matrices) are different in order to extract features in multiple aspects.

The third layer is the pooling layer. In this paper, the maximum pooling method is adopted, that is, taking the largest element max{c} in the convolution layer vector c obtained by convolution of each convolution window as the eigenvalue, thereby extracting the eigenvalue p_j corresponding to each convolution window (j= 1, 2, ..., w, where w is the total number of convolution windows), and all the eigenvalues p_j are sequentially spliced to form a vector p$\in R^w$ of the pooling layer, p is a vector representing the global features of the sentence. The pooling process not only achieves further extraction of features, but also reduces the dimension of features and improves classification efficiency.

The fourth layer is the output layer. The output layer is fully connected to the pooling layer, and the pooling layer is input to p, and the vector p is classified by the softmax classifier, and the final classification result is output.

4 Experimental Results and Analysis

4.1 Experimental Environment and Data

The experimental environment of this paper is Window10 system, the experimental programming language is Python3.0, the development tool is PyCharm, and the deep learning Tensorflow framework is used. The experimental data set is the user's electricity text data provided by the State Grid, with a total of 100,000 pieces of data. The data is divided into training sets and test sets to evaluate the classification effect of the model.

4.2 Experimental Design

This paper uses the two-layer framework model of Stacking ensemble to classify power text data. Because the CNN model is used in the second layer, the classification results are measured by precision and recall (Recall) and F1 values.

(1) In order to explain the algorithm model of this paper to the traditional model, the classification results of the Stacking ensemble two-layer framework model used in this paper are compared with the classical CNN model classification results. In addition, in order to illustrate the advantages of the S-TFIWF algorithm classification based on the Boolean model, the text is classified using the traditional TFIDF algorithm on the same data set, and then the classification results are compared. In order to eliminate the incomparability of the experimental results due to the different ways of constructing the features, the feature construction method using traditional machine learning is also based on the word vector, and the feature of each power text clerk is taken as the mean of all word vectors.

(2) In order to illustrate the influence of the selection ensemble in the two-layer framework model of Stacking ensemble on the results, the values of different Kappa coefficients k_i are designed, and the classification effects are compared under different k_i values.

4.3 Experimental Results and Analysis

(1) Table 1 shows the comparison results of the overall average Precision value, Recall value, and F_1 value of different classification models.

Table 1. Comparison of average classification results for different classification models

Classification model	Precision	Recall	F_1
Traditional TFIDF algorithm model	0.9672	0.9673	0.9672
Traditional CNN model	0.9781	0.9780	0.9780
The model used in this paper	0.9870	0.9870	0.9870

It can be seen from Table 1 that for the same data, the classification effect of the S-TFIWF algorithm model based on the Boolean model is more accurate and better than the traditional TFIDF algorithm model. The Boolean model-based S-TFIWF algorithm Stacking selectively integrates the two-layer framework model to achieve more accurate classification results than the traditional CNN model.

The comparison of these two models shows that the model in this paper performs better in the classification of text than the traditional classification algorithm in the three aspects of Precision value, Recall value and F_1 value. It also shows that the framework model of this paper introduces Boolean. Model S-TFIWF algorithm, Stacking selectively integrates the advantages of the two-layer framework model. Because the different test sets are selected to test the framework model of this paper, it also shows that the proposed algorithm has better generalization ability.

The classification algorithm in this paper is compared with that in reference [4], as shown in Fig. 4, and the classification accuracy of the arithmetic in the figure is compared. The classification accuracy of this algorithm is about 0.983, and that of the literature is about 0.976. It is proved that the framework of selective integration has been improved to some extent in text classification.

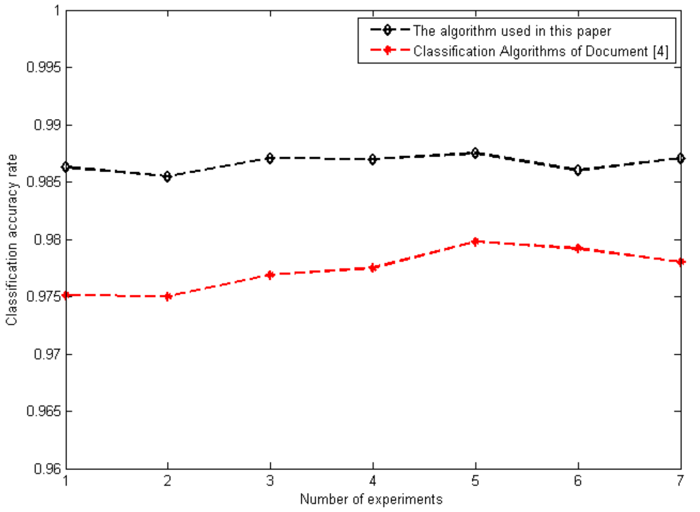

Fig. 4. Comparisons between the algorithms in this paper and those in literature

(2) Table 2 shows the effect of different Kappa coefficients k_i on the overall average Precision value, Recall value, and F_1 value.

Table 2. Comparison of average classification results for different k_i values

k_i Value	Precision	Recall	F_1
k_i =0.6	0.9753	0.9750	0.9750
k_i =0.7	0.9791	0.9790	0.9790
k_i =0.8	0.9870	0.9870	0.9870
k_i =0.9	0.9821	0.9820	0.9820

It can be seen from Table 2 that when the first layer of the framework model is selected and ensemble, the influence of different Kappa coefficients k_i on the experimental results is also very different. When the value of k_i is less than 0.8, the algorithm model the effect is enhanced with the increase of k_i. When the value of k_i is greater than 0.8, the effect of the algorithm model will also decrease. Therefore, when k_i is 0.8, the classification effect of the algorithm model is the best.

In Fig. 5, it can be clearly seen that when the k_i is taken as 0.8, the values of the three indicators of the model reach the extreme point, and the values on both sides of the 0.8 are reduced, therefore, when k_i takes 0.8, the integration effect of each classifier is optimal. The comparison of the values of k_i shows that the ensemble strategy of each base classifier determines the classification effect of the model when performing selective ensemble. The experimental results show that the model framework of this paper achieves the best 0.8 classification effect when performing selective ensemble.

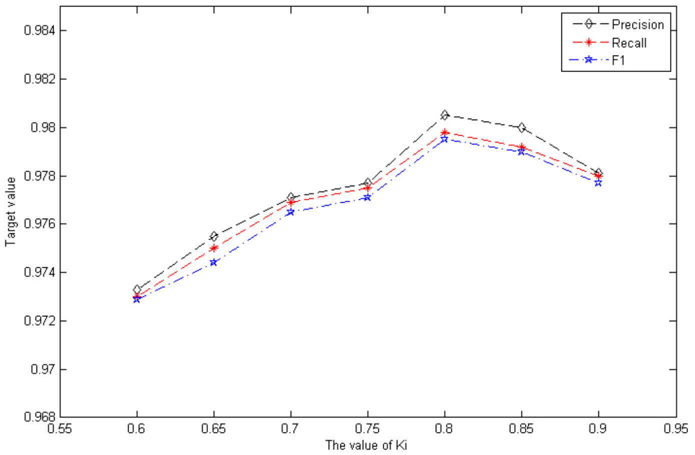

Fig. 5. Comparison of average classification results for different k_i values

5 Summary

In this paper, the S-TFIWF based on Boolean model is used to process the weight of all power text data, and then processed by Stacking's selectively ensemble two-layer framework model. The word vector obtained from a large amount of text information in the second layer, expressed as a feature of the text. Compared with the traditional machine learning method of extracting features, word2vec can automatically condense the semantic information into the mathematical vector, and splicing the word vector with the characteristics of the first data in the second layer, using the vector as the input feature of the classification model. Based on the classic CNN classification, this paper introduces the idea of selective ensemble, which makes the expression of features more abundant, and makes the generalization ability of the model stronger, so that the CNN model has a certain degree of improvement on the classification effect of text.

References

[1] Y.Q. Song, G.L. Zhou, Y.L. Zhu, Status and challenges of large data processing technology for smart grid. Power grid technology 37(4)(2013) 927-935.

[2] C. Zhao, Semi-supervised Chinese text categorization based on selective ensemble, [dissertation] Hangzhou: Zhejiang Industrial and Commercial University, 2018.

[3] N. Jiang, H.Y. Yang, Q.C. Gu, J.Y. Huang, Machine learning and its algorithms and development analysis, Information and Computer (Theoretical Edition) 1(2019) 83-84+87.

[4] Q.P. Zhou, C.G. Tan, H.J. Wang, Z.X. Zhan, An improved KNN text classification algorithm based on clustering, Computer Application Research 33(11)(2016) 3374-3377+3382.

[5] Z.G. Jin, J. Han, Q. Zhu, An emotional analysis model combining deep learning and integrated learning, Journal of Harbin University of Technology 50(11)(2018) 32-39.

[6] W.B. Pan, G. Cheng, X.J. Guo, Y. Wang, Embedded network flow feature selection based on selective integration strategy, Journal of Computer Science 37(10)(2014) 2128-2138.

[7] M.Q. Zhang, F.Q. Di, J. Liu. General Steganalysis based on selective integrated classifier, Journal of Sichuan University (Engineering Science Edition) 47(1)(2015) 36-41.

[8] Q.L. Zhao, Y.H. Jiang, M. Xu, Classification and comparison of selective integration algorithms, Computer Engineering and Science 34(2)(2012) 134-138.

[9] M. Tang, L. Zhu, X.C. Zou, A document vector representation based on Word2Vec, Computer Science 43(6)(2016) 214-217+269.

[10] K.D. He, Z.T. Zhu, Y. Cheng, Text classification method based on improved TF-IDF algorithm, Journal of Guangdong University of Technology 33(5)(2016) 49-53.

[11] Y. Kim, Convolutional neural networks for sentence classification, in: Proc. the 2014 Conference on Empirical Methods in Natural Language Classification (EMNLP 2014), 2014.

[12] Z.Q. Liu, H.F. Wang, J. Cao, J. Qiu. Research on text classification model of power equipment defects based on convolutional neural network, Power Grid Technology 42(2)(2018) 644-651.

[13] R.B. Sun, Quenching of solutions of Quasilinear parabolic systems with singular terms under second boundary conditions, Journal of Central South University for Nationalities (Natural Science Edition) 37(2)(2018) 133-136.