Text Classification Feature Extraction Method Based on Deep Learning for Unbalanced Data Sets

Li Lin[1(✉)] and Shu-xin Guo[2]

[1] School of Computer Engineering, Jimei University, Xiamen 361021, China
xd220210@163.com
[2] Jilin University of Finance and Economics, Changchun 130117, China

Abstract. In order to fully realize the classified search of text data information, a text classification feature extraction method for imbalanced data sets based on deep learning is proposed. With the help of trestle automatic encoder and depth confidence network, the preliminary definition of text semantic category conditions is completed, and the text semantic classification processing based on depth learning algorithm is realized. On this basis, pre-processing and debugging of text parameters are implemented, and the dimensionality reduction standards related to the text features of the data set to be extracted are established through the expression of the characteristic behavior. The experimental results show that with the application of the new classification feature extraction method, the number of correctly classified documents starts to increase substantially, which meets the practical application requirements for the classification and search of text data information.

Keywords: Deep learning · Unbalanced data sets · Text features · Classification and extraction

1 Introduction

Deep learning is a new research direction in the field of machine learning. It is introduced into machine learning to make it closer to the original goal artificial intelligence. Deep learning is to learn the inherent laws and representation levels of sample data. The information obtained during these learning processes is of great help to the interpretation of data such as text, images, and sound. Its ultimate goal is to enable the machine to be able to analyze and learn like human beings, and to recognize data such as words, images and sounds. Deep learning is a complex machine learning algorithm, which has achieved much better results in speech and image recognition than previous related technologies. Deep learning has made many achievements in search technology, data mining, machine learning, machine translation, natural language processing, multimedia learning, voice, recommendation and personalized technology [1, 2]. Deep learning enables machines to imitate human activities such as audio-visual and thinking, solves many complex pattern recognition problems, and makes great progress in AI related technologies. Deep learning is a kind of machine learning, and machine

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 320–331, 2021.
https://doi.org/10.1007/978-3-030-67871-5_29

learning is the only way to realize artificial intelligence. The concept of deep learning stems from the study of artificial neural networks. A multi-layer perceptron with multiple hidden layers is a deep learning structure. Deep learning combines low-level features to form more abstract high-level representation attribute categories or features to discover the distributed feature representation of data. The motivation for studying deep learning is to build a neural network that simulates the human brain for analysis and learning. It mimics the mechanism of the human brain to interpret data, such as images, sounds, and text.

Text classification uses computers to automatically classify and mark text sets (or other entities or objects) according to a certain classification system or standard. According to a set of labeled training documents, it finds the relationship model between document features and document categories, and then uses the relationship model to judge the new document categories. Text classification has gradually changed from knowledge-based method to statistical and machine learning based method [3]. Text classification generally includes text expression, classifier selection and training, classification result evaluation and feedback, etc. The text expression can be subdivided into text preprocessing, indexing and statistics, feature extraction and other steps. The overall function module of the text classification system is: preprocessing: formatting the original corpus into the same format for subsequent unified processing; indexing: decomposing the document into basic processing units, while reducing the cost of subsequent processing; statistics: word frequency statistics, the correlation probability of items (words, concepts) and classification; feature extraction: extracting the characteristics reflecting the document theme from the document Feature; classifier: training of classifier; evaluation: analysis of test results of classifier.

Therefore, this paper proposes a deep learning based text classification feature extraction method for imbalanced data sets. Firstly, this paper introduces the research status and significance of text classification, and expounds the definition, method and process of text classification. Secondly, the text classification algorithm is described, in which KNN and SVM classification algorithm are introduced in detail. The common feature selection methods are introduced and their advantages and disadvantages are analyzed. Thirdly, on the basis of the above, one of the feature selection methods CHL statistical method is improved. Finally, the improved method is verified on the basis of experiments, which shows the feasibility of the improved method.

2 Text Semantic Classification Based on Deep Learning

Text semantic classification is the basic processing link of the application of text classification feature extraction method in unbalanced data set. Under the support of deep learning algorithm, the specific operation process is as follows.

2.1 Stacked Automatic Encoder

The trestle type automatic encoder is a neural network model with multiple hidden layer neurons. A typical trestle type automatic encoder structure is shown in Fig. 1.

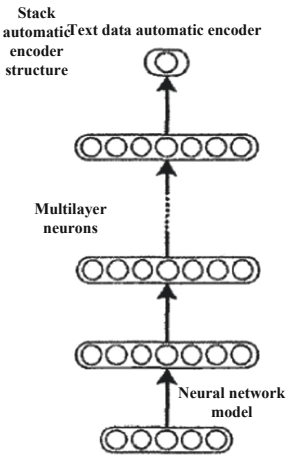

Fig. 1. Structure diagram of the stack type automatic encoder

The trestle type automatic encoder can be regarded as a combination of multiple text data automatic encoders. With the support of the deep learning algorithm, in order to achieve the learning speed of the text data in the unbalanced data set, and to initialize the weight value to a better eigenvalue state, in order to obtain the local optimal value close to the global optimal value, the text classification features to be encoded must be pre trained. In the process of pre training, starting from the input layer, the adjacent two layers are regarded as a separate set of restricted sampling samples for training, and the output of the lower layer is regarded as the input of the higher layer, so as to complete the initialization of the weight [4]. In the process of training, since the deep learning algorithm still has the problem of text gradient dissipation, the unbalanced data set is expanded. At this time, the process of bottom-up propagation can be regarded as encoding the input data, while the process of downward propagation can be regarded as the process of decoding. At the same time, in order to prevent the classification features from overfitting the data and affecting the promotion ability, a certain amount of noise will be added to the weights during fine-tuning.

2.2 Deep Confidence Network

The deep belief network is a directed graph model. With the support of a stack autoencoder, you can specify the classification feature structure of the text information in the query imbalanced data set. The weights between variables derive the state of the hidden variables as shown in Fig. 2.

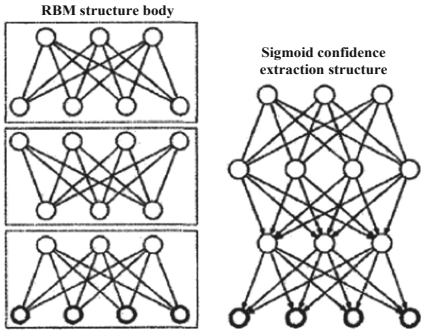

Fig. 2. Structure of depth confidence network

According to Fig. 2, the top two layers of deep learning mechanism can be regarded as an RBM structure subject, while the network below the top layer can be regarded as a directed sigmoid confidence extraction structure. One of the problems of deep confidence networks is how to initialize the feature weights of the text of the data set. It is usually very difficult to optimize the nonlinear deep network with multiple hidden layers, and only the initialization can get better weights, and the network can avoid Fall into a local minimum. The feature extraction method of information initialization based on deep learning mechanism regards two adjacent layers as RBM structure, and trains the network from bottom to top [5, 6]. If the classification characteristics of text data do not change, after unsupervised pre training, the whole network uses supervised learning to fine tune, and finally get the trained DBN data set information. The process of DBN information fine-tuning is similar to that of depth auto encoder, which is also fine-tuning after expansion.

2.3 Text Semantic Category Definition

Text semantic classification, as an important means of data set feature analysis, is widely used in information search, information annotation and other occasions. In some professional fields, unbalanced data set classification has achieved high accuracy, such as big data information recognition, handwritten digit recognition. However, the classification of text features is still a challenging issue due to the reasons for changes in transmission traffic, network transportation rate, data set flexibility, and the complexity of the data set content itself. The semantic-based text classification method starts from the imbalanced data set itself, uses a specific feature extraction method to extract the semantic features in the data, and gradually builds a semantic hierarchy of information on this basis, and finally uses the learned high-level information semantic features sort. The semantic features of data sets are hierarchical, among which the low-level information features are of low abstraction, high correlation with the information data content itself, high-level information features are of high abstraction, and low correlation with the information data content. Therefore, it may be considered to establish a hierarchical learning model, and use unsupervised learning to learn information data to obtain the characteristics of the data itself. As the level increases, the

learned features become a higher-order feature description of the input data. Using deep learning model to learn the semantic features of information data, through the increase of depth to improve the abstraction of the model, so as to establish the semantic hierarchy, and then using the classification model to realize the semantic based classification of data sets. Let $s_{\max}$ represent the high-level classification features of the input information of the dataset, $u_{\max}$ represent the classification hierarchy features of the semantics of the highest-level dataset, and $u_{\min}$ represent the classification hierarchy features of the semantics of the lowest-level dataset. The category definition expression of text semantics is:

$$R = \frac{s_{\max}}{2}\sqrt{\frac{y(u_{\max} - u_{\min})^{\varepsilon}}{p^2 - q^2}} \tag{1}$$

Among them, y represents the average transport rate of the unbalanced text data information in the deep learning network, p represents the text feature parameter of the data set information itself, and q represents the text feature parameter in the given semantic model.

3 Text Classification Feature Extraction from Unbalanced Dataset

With the support of the deep learning principle of text semantic classification, according to the application process of text preprocessing, text feature representation, and dimensionality reduction of features to be extracted, the smooth application of text classification feature extraction method for imbalanced data sets is completed.

3.1 Text Preprocessing

From the perspective of the text classification process, regardless of whether Chinese text classification or English text classification is performed, the text used must be preprocessed to remove some useless information. This is to reduce the complexity and complexity of the next steps. The burden of calculation. First, we need to segment words (also known as segmentation). If the text to be classified is Chinese, word segmentation is an essential step, which is to divide the consecutive sentences into individual words (because there is no obvious segmentation mark between a word and a word in Chinese, and the word is the smallest unit in the text that can be used independently). For example, the sentence "I am a student" should be divided into "I/Yes/One/Student". However, if the text to be classified is English, there is no need for word segmentation because the space and punctuation in English have already played a role in word segmentation. The quality of word segmentation has a great impact on the performance of the whole text classification system, which is mainly because the text information used in the later process is all preprocessed text. If the word segmentation is not done well, the whole training text set will be inaccurate, resulting in the accuracy rate of the classification is reduced.

In terms of current development, word segmentation methods include dictionary-based methods, statistical-based methods, and hybrid methods. Among them, the dictionary-based method generally needs to meet the three conditions of the word segmentation dictionary, the order of scanned text, and the matching principle. The order of scanning text generally has three scanning methods: forward, reverse, and bidirectional. The matching principles generally include three methods: forward matching, reverse matching, and bidirectional matching. The word segmentation method based on statistics uses word and word occurrence probability as the basis of word segmentation [7, 8]. The common methods include hidden Markov model, n-ary grammar model and maximum entropy model. The mixed method is to combine two or more methods to segment words. There are two standards of word segmentation: segmentation speed and segmentation precision. Obviously, the accuracy of segmentation is the most important, because it directly determines the quality of classification. In terms of current development, the main limitation of segmentation is that segmentation efficiency is low and the effect of disambiguation is not good.

3.2 Text Feature Representation

At present, there are vector space model, Boolean model and probability retrieval model.

(1) Vector space model

The vector space model defines the correlation between documents as a similarity between them. It believes that the more similar a document is to a user query, the more relevant it is to a user query. Deep belief networks represent documents as a vector in a high-dimensional word space. Each dimension in the vector represents the weight of the corresponding feature in the document. The measure of similarity is the angle cosine.

(2) Boolean model

The Boolean model is a special case of the vector space model. The classic Boolean model can only be used to calculate the correlation between user queries and documents in information retrieval, but it cannot use the model to calculate the deeper similarity of two documents and cannot be used in more text processing. At present, scholars have proposed a variety of extended Boolean models, so that the correlation is no longer simply 0 and 1, but becomes a number between [0, 1].

(3) Probability retrieval model

Both the Boolean model and the vector space model treat the document representation terms as independent items, ignoring the correlation between the representation terms. The probabilistic model takes into account the internal relations between terms and documents, and uses the probability dependence between the terms and between the terms and the documents to retrieve information.

The computer does not have human intelligence. After reading an article, people can have a vague understanding of the content of the article according to their own understanding ability. Fundamentally, the computer can only know 0 and 1. So, like all

machine learning problems, if you want the computer to automatically classify text, you need to express the text as a feature, such as words, words, n-grams, phrases, concepts, etc. Obviously, this will lose a lot of information about the content of the article, but this representation can formalize the processing of the text and can achieve better results in text classification.

(a) word

Word is the most widely used feature in text classification. For a document, the most intuitive way is to use words and phrases as the characteristics of the text. For English articles, each word has been separated by a space, and the feature words can be obtained directly. However, due to the changes of word form in English, such as the singular and plural numbers of nouns, the tense changes of verbs, the prefixes and suffixes of words, a process of stem extraction is needed. For Chinese, because there is no pause between words, we need to use dictionaries and special word segmentation techniques.

(b) N-gram items

N-gram is a text representation independent of language. Because the n-gram string method does not consider whether the semantic unit of the text is word, word or phrase at all, but regards the whole text as a string composed of different characters, so it can conveniently represent all kinds of language documents including Chinese and Arabic. For Chinese, n-gram items are generally composed of adjacent words. For English, n-gram can be composed of adjacent words or letters. The N-gram item is a feature of the document, which can avoid huge dictionaries and complicated word segmentation programs. In general, using the same classification method, the effect of word-based text classification is no better than that based on N-gram items. In the case of a small number of features, the classification effect based on N-gram items is better than that based on words. The obvious feature of the N-gram method is the large amount of calculation, so there are not many applications.

(c) Phrase

Phrase notation is widely used in the field of text classification. This notation improves the semantic content of feature vectors and restores some useful information thrown away by word notation. But its expressive power is not obvious. This representation method reduces the statistical quality of feature vectors and makes them more sparse, which makes it difficult for machine learning method to extract statistical characteristics for classification.

(d) Concept

The concept has a higher abstraction. A concept can correspond to a word in a text, or it can correspond to several semantically related words. Using concept space can greatly reduce the dimension of feature space, thus reducing the training time of classifier and the time for similarity comparison. Therefore, concept-based text classification is based on word-based classification in terms of time efficiency; at the same time, because a concept can merge multiple keywords with synonymous relationships, it can avoid an important classification feature due to the dispersion of key times and

weaken The weight of the classification; again, mapping one keyword to multiple concepts can avoid the feature ambiguity caused by using only keywords as features. However, because of the complexity of the concept, it will cost unimaginable human and material resources. Secondly, the establishment of concept depends on experts or domain experts, so it has strong subjectivity. Therefore, in the practical application, the concept based application is not ideal.

3.3 Dimensionality Reduction of Features to Be Extracted

Text feature dimensionality reduction is a key step in text classification. An important problem in text mining is the existence of high-dimensional feature space, which is composed of words or phrases in the text. Many traditional methods are difficult to deal with. High-dimensional feature sets are not necessarily all important and beneficial to machine learning, but also increase the burden of machine learning. Without affecting the accuracy of feature classification, it is necessary to reduce the number of high-dimensional features in the text description space. This process is feature dimensionality reduction [9, 10].

The traditional text feature dimensionality reduction method uses a single evaluation function to calculate the feature weight [11–13]. Because the traditional feature evaluation function only pays attention to the single aspect of weight calculation and ignores other important factors, the traditional feature dimension reduction method is not effective. In the process of text mining, it is necessary to consider the influence of multiple factors, effectively combine feature extraction and feature selection, and jointly reduce the feature dimension. The combined feature dimensionality reduction method presented in this paper not only combines the feature items with similar contributions to the classification category into new feature items, but this feature extraction operation has the method of using attribute reduction on the merged feature items. Selected feature selection operation. In this paper, we first use the pattern aggregation theory to fuse the features that have similar contribution to the classification, and then use the method of rough set decision table to connect feature selection with text classification, that is, attribute reduction is carried out according to the importance of features in classification. This method not only reduces the complexity of training, greatly reduces the dimension of feature, but also improves the accuracy of feature dimension reduction.

4 Practical Ability Testing

4.1 Experimental Data and Steps

In order to verify the practical application value of text classification feature extraction method based on deep learning unbalanced data set, the following practical detection experiments are designed. Before text classification, the first step is to prepare training text set and test text set. All text set data used in this experiment are from http://www.nlp.org.cn/ website, from which ten categories of computer, environment, transportation, economy, sports, medicine, education, art, politics and military are extracted for

experiment. There are 1887 training document sets and 934 test document sets. The distribution of categories is shown in Tables 1 and 2.

Table 1. Distribution of training concentration categories

Class alias	Computer	Traffic	Economic	Sports	Medicine
Number of training	135	145	215	301	135
Class alias	Surroundings	Education	Art	Political	Military
Number of training	136	150	165	340	165

Table 2. Distribution of test set categories

Category name	Computer	Traffic	Economic	Physical education	Medicine
Number of tests	65	70	108	150	70
Category name	Surroundings	Education	Art	Political	Military
Number of tests	66	73	82	167	83

Taking the text feature classification of unbalanced data sets as an example, the whole experimental process mainly includes the following steps:

First, the selected training text set is segmented.

Secondly, SVM algorithm and KNN algorithm are used to train the segmentation results. Some parameters should be selected during training. For example, feature selection, feature dimension and weight calculation function. These parameters will affect the result of classification. In this process, we mainly test the improved feature selection method.

In the training process of this experiment, the linear kernel function is selected for the kernel function of SVM. The other parameter selections are the same in KNN and SVM, respectively, the feature selection method selects CHI statistical method; the feature dimension selects 1000; the weight calculation selects the established support function.

Third, a classification model (classifier) is formed on the basis of the above.

Fourth, set the parameters of the new test text set. If KNN is used for classification, select 35 for K value, and then classify.

Fifth, after the classification is completed, view the classification results. Take the result of art.

4.2 Number of Documents

This experiment mainly uses the evaluation methods of recall rate (T) and precision rate (P) to evaluate the documents in the test document set, and T and P are expressed by formula 2 and formula 3. In addition to T and P, this experiment also used histogram and confusion matrix to evaluate it.

$$T = \frac{\text{Number of documents classified correctly in a category}}{\text{Total number of documents in this category}} \times 100\% \quad (2)$$

$$P = \frac{\text{Number of documents classified correctly in a category}}{\text{Number of documents assigned to this category}} \times 100\% \quad (3)$$

Experimental results 1: the experimental results are analyzed from the number of documents classified by each category. Table 3 compares the number of classified documents before and after improvement.

Table 3. Comparison of the number of documents after KNN classification

Category	The number of documents classified as belonging to this category			The number of correctly classified documents in documents belonging to a certain category			The total number of documents belonging to this category
	Tradition CHI	Improve CHI_1	Improve CHI_2	Tradition CHI	Improve CHI_1	Improve CHI_2	
Traffic	63	65	65	61	63	63	70
Surroundings	57	64	58	52	54	52	66
Computer	58	57	57	57	56	56	65
Education	72	69	68	66	65	65	73
Economic	127	124	121	101	103	102	108
Art	76	80	82	73	77	76	82
Physical education	159	154	154	149	146	147	150
Political	199	199	194	157	161	159	167
Medicine	67	64	69	65	63	65	70
Military	62	58	60	53	53	53	83

It can be seen from Table 3 that the number of documents correctly classified in the improved method is better than that in the traditional method, and the number of some categories is lower than that in the traditional method, for example, in the improved chic method, the number of documents correctly classified in the documents belonging to the categories of computer, education, sports and medicine; in the improved chic method, the number of documents belonging to the counting the number of documents correctly classified in computer, education and sports documents. This may happen because the selected test document appears too little in the training document set. But from the overall classification results, the number of correctly classified documents has increased.

4.3 Recovery and Precision Rates

Experimental result 2: the experimental results are analyzed from the recall and precision of each category. From the number of document classification in Table 3, the recall and precision of each category can be calculated. Take the traffic categories in the traditional Chi method as an example, $T = \frac{61}{70} \times 100\% = 87.143\%$, $P = \frac{61}{63} \times 100\% = 96.825\%$,Table 4 compares the recall rate and precision rate before and after improvement.

Table 4. Comparison of recall rate and recall rate after KNN classification

Category traffic	Tradition CHI		Improvement CHI_1		Improvement CHI_2	
	T	*P*	*T*	*P*	*T*	*P*
Environmental science	87.143%	96.825%	90.000%	96.923%	90.000%	96.923%
Computer	78.788%	91.228%	81.818%	84.375%	78.788%	89.655%
Education	87.692%	98.276%	86.154%	98.246%	86.154%	98.246%
Economics	90.411%	91.667%	89.041%	94.203%	89.041%	95.588%
Art	93.519%	79.5285%	95.370%	83.065%	94.444%	84.298%
Sports	89.024%	96.053%	93.902%	96.250%	92.683%	92.683%
Politics	97.333%	93.711%	97.333%	94.805%	98.000%	95.455%
Medicine	94.021%	78.894%	96.407%	80.905%	95.210%	81.959%
Military	92.857%	97.015%	90.000%	98.438%	92.857%	94.203%
Category	63.855%	85.484%	63.855%	91.379%	63.855%	88.333%

It can be seen from Table 4 that the improved classification effect is better than the traditional method as a whole, and the recall rate and precision rate of some categories are lower than the traditional method. For example, in the improved CHIC method, the computer, the recall rate of education and medicine category, the precision rate of environment and computer category; in the improvement CHI: method, the recall rate of computer and education category, the precision rate of environment, computer, art and medicine category. But from the overall classification results, the recall rate and precision rate have been improved.

5 Conclusion

With the development of the network, a lot of information appears on the network. How to find the information we need quickly from the network has become more and more important [14, 15]. Text classification is one of the methods to solve this problem. Therefore, this paper proposes a deep learning based text classification feature extraction method for imbalanced data sets, and defines the text semantic categories through the trestle automatic encoder and deep confidence network. According to the application process of text preprocessing, text feature representation and feature dimension reduction, the smooth application of text classification feature extraction

method is completed. The experimental results show that the number of documents correctly classified by this method is more, and the recall rate and recall rate of KNN are improved.

References

1. Chen, W., Liu, X., Lu, M.: Feature extraction of deep topic model for multi-label text classification. Pattern Recogn. Artif. Intell. **32**(9), 785–792 (2019)
2. Wang, Y., He, Y., Zou, H., et al.: WordNG-Vec: a word vector model applied to CNN text classification. J. Chin. Comput. Syst. **40**(03), 37–40 (2019)
3. Song, C., Chen, X., Niu, Q.: Improved feature selection method based on CHI for text categorization. Microelectron. Comput. **35**(09), 80–84 (2018)
4. Han, D., Wang, C., Xiao, M.: Multi-label text classification method based on rotating forest and AdaBoost classifier. Appl. Res. Comput. **35**(12), 141–144 (2018)
5. Yin, Y., Yang, W., Yang, H., et al.: KNN text classification algorithm based on search improvement. Comput. Eng. Des. **39**(09), 231–236 (2018)
6. Tong, X., Guo, P., Xu, P., et al.: Fusing hyperspectral features and image deep features for classification and retrieval of meat. Sci. Technol. Food Ind. **39**(23), 261–266+272 (2018)
7. Xuan, Q., Fang, B., Wang, J., et al.: Pearl multi-feature classification method based on support vector machine. J. Zhejiang Univ. Technol. **46**(05), 5–12 (2018)
8. Hua, S., Hu, S., Gao, L., et al.: Research on fish counting and species recognition system of fishway based on image feature extraction. Water Power **44**(12), 90–9 +128 (2018)
9. Lv, W., Deng, W., Chu, J., et al.: Arrhythmia classification based on feature selection method of S-transform. J. Data Acquis. Process. **33**(2), 306–316 (2018)
10. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018). https://doi.org/10.1155/2018/2016976
11. Liu, S., Yang, G. (eds.): ADHIP 2018. LNICST, vol. 279. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-19086-6
12. Sridharan, K., Sivakumar, P.: A systematic review on techniques of feature selection and classification for text mining. Int. J. Bus. Inf. Syst. **28**(4), 504–518 (2018)
13. Ferreira, C.H.P., De Franca, F.O., Medeiros, D.R.: Combining multiple views from a distance based feature extraction for text classification. In: IEEE Congress on Evolutionary Computation, pp. 1–8. IEEE (2018)
14. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mobile Netw. Appl. **24**(1), 1–4 (2019)
15. Saikia, L.P., Singh, S.: Feature extraction and performance measure of requirement engineering (RE) document using text classification technique, pp. 1–6 (2018)