ATLANTIS PRESS

Advances in Computer Science Research, volume 83

8th International Conference on Social Network, Communication and Education (SNCE 2018)

Research on Text Classification Method Based on Multi-type Classifier Fusion

Zeng Meilin

(Jiangxi Industrial and Trade Vocational and Technical College 330038)

Keywords: Text classification; Classifier fusion; Principal component analysis; Latent semantic index

Abstract. Most of the traditional text classification methods use a single classifier, but different classifiers have different emphasis on classification tasks, which makes a single classification method have some limitations. This paper presents a text classification method based on multi-type classifier fusion, which uses word2vecand principal component analysis (PCA) as feature extraction method for multi-type classifier fusion. At the same time, the problem of category information is ignored in the weighted voting method of multi-type classifier, and the method of classifier weight calculation is adopted. The experimental results show that the multi-type classifier fusion method is in binary. Good performance has been obtained on corpus, multivariate corpus and specific corpus. The classifier weight calculation method with class weighting improves the classification performance by 1.19% compared with multi-type classifier fusion method.

China Library classification number: TP391 Document identification code: a national standard subject classification code: 520.4070

Introduction

The text classification method in machine learning is the key technology[1~2]to process and manage the document data. The text classification method has been widely studied at home and abroad. For example, it is simple to use K nearest neighbor KNN method. Text spam short message classification [3]. Goudjil et al. [4] select samples by using the posterior probability provided by SVM classifier, and classify them by using selected samples.

The above research methods have achieved very good classification effect. However, the K value in the KNN method is manually set and has great objectivity. How to determine the kernel function of high-dimensional space in SVM is one of the difficulties at present. The above methods all use a single classifier to classify the texts, and the text data involves a very wide field, which makes a single classifier not well cover more fields. Therefore, this paper introduces a multi-classifier fusion text classification method.

Text categorization is the distribution of documents with similar textual content into one or more predefined categories, and feature extraction methods have an important role in improving classifier performance[6]. For example, in the document[7] the use of kinetic energy theorem and TFIDF feature extraction method to solve the microblogging subject detection problem. In the document[8], word2vec is used as an automatic feature extraction tool, and then sentence vectors are used to complete the classification.

The above feature extraction methods have achieved good results in their respective classification tasks. However, the TFIDF method only considers statistical indicators of words and does not consider the semantic knowledge of characteristic words. Word2vec does not consider the statistical characteristics of feature words. The above method is a single feature extraction method and the feature words of each feature extraction method are different, so the above methods have certain limitations. Therefore, in order to better represent the features of the text, this article uses a fusion feature extraction method.

In summary, for the single problem of text classifier and feature extraction methods, this paper proposes a text classification method that combines multiple types of classifiers. At the same time, the classifier-weighted voting decision method does not consider the problem that the classifier has different contributions to each class, and proposes a category-weighted classifier weight calculation

method.

Multi-type Classifier Fusion Text Classification

The method of multi-category classification fusion is to make the feature words in the feature space vector richer by fusing different feature extraction methods, and to enrich the expression form of the text by fusing multiple feature extraction methods, and then classify the text by using a classifier. The multi-category classification fusion method in this paper includes the following feature extraction methods: word2vec, TF-IDF, LDA, and LSI.

Word2vec

In 2013, Mikolov et al. proposed a Word2vec open source software. The Word2vec method converts words into word vectors through neural network methods. In the process of training the word vectors, the word generation word table in the training data set is first extracted by using CBOW or Skip-Gram model to come to each word word vector, the model diagram shown in Figure 1.

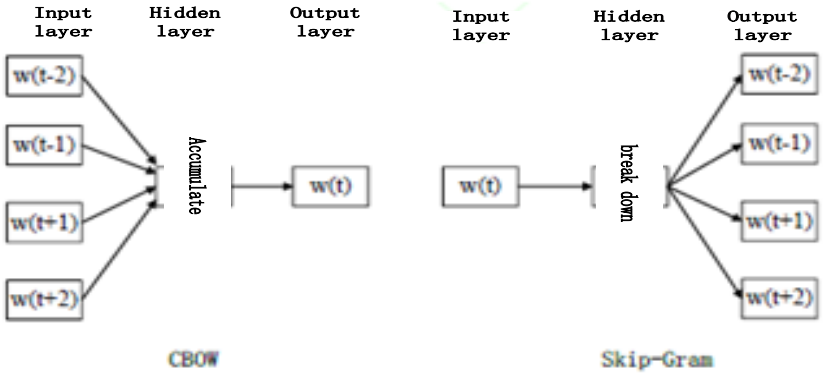

Figure 1 CBOW and Skip-Gram models

As shown in Figure 1, the CBOW and Skip-Gram models are a reverse process. The CBOW model uses the t words before and after the word to be predicted to predict the current word, and the kip-Gram model uses the current prediction word to predict each of the t words before and after.

TF-IDF

TF-IDF is a classical feature weight calculation method. TF-IDF consists of TF (word frequency) and IDF (inverse document frequency). The formula is as follows:

$$tifidf(w) = tf(w) \times idf(w) \tag{1}$$

Where: tf(w) is the number of occurrences of word w in the text, idf(w) is the inverse document frequency of word w, and idf(w) is calculated as shown in equation (2).

$$idf(w) = \log \frac{A}{B(w)} \tag{2}$$

In which: A represents the total number of texts in the training set, and B(w) represents the number of files containing the word w.

LDA

LDA is a topic model and is a three-layer Bayesian model of word-document-subject. The subject model trains through the training set to derive the Dirichlet distribution of the topic and the polynomial distribution function between the topic and the word. The method first determines a topic and then selects one word in the topic until all words are traversed.

LSI

LSI is an unsupervised data mining technology that has a good effect on semantic issues such as polysemy. In the latent semantic indexing method, the singular value decomposition method is used to decompose the feature vector space to achieve the purpose of dimension reduction.

Multi-type Classifier Fusion

Through the feature extraction method in Section 1.1, four sets of different feature vector spaces are generated. The first is the vector space of word2vec generated by the CBOW method, the second is the vector space generated using TFIDF, the third is the semantic vector space generated by LSI, and the last one is the LDA vector space generated using the theme model. The method of multi-type classifier fusion is to use feature extraction methods to generate complementarity between vector spaces. The multi-class classifier method model is shown in Fig. 2. The numbers in the triangles are the weights of the classifiers.

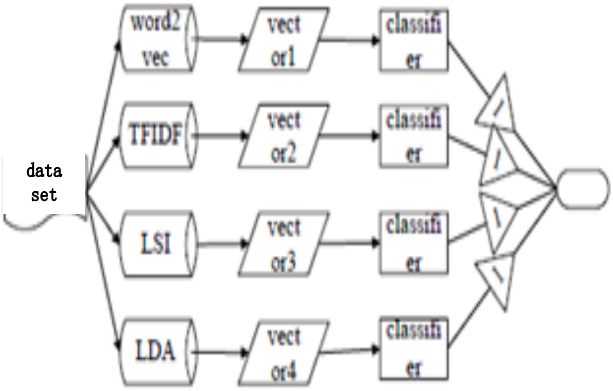

Figure 2. Multi-class classifier fusion diagram

Category-weighted Multi-type Classifier Fusion

Multi-classifier fusion can use different classifiers to complete different tasks, so as to avoid considering incomplete problems. Different classifiers have different classification capabilities for the same sample. Therefore, each classifier has different contribution capabilities for each sample. The classifier-weighted voting method is one of the methods of classifying voter voting. Classification performance (the correct recognition rate of trained samples for training samples) has the following advantages as a classifier-weighted voting method. Different classifiers have different recognition rates for the same sample. When combining classifiers to make classification decisions, the classification results tend to be different. With classifiers with good classification performance, the best decision-making performance is achieved. Therefore, this paper uses the classification performance as the weight of the classifier. The classification performance formula is shown in formula (3)(4).

$$\varepsilon = \frac{errorNum}{textNum} \tag{3}$$

$$\alpha = in(\frac{1-\varepsilon}{\varepsilon}) \tag{4}$$

In which: errorNum is the number of samples that the classifier does not correctly classify; textNum is the total number of samples in the sample data set; it is the weight of the classifier to the data set.

Figure 3 is a schematic diagram of a traditional binary classification sample. There are a total of 100 data points in the figure. There are 40 training data points in each category. Each category has 10 test data points, which represents training data category 1, and x represents Training data category 2, Δ represents test data in test data category 2. This section uses KNN and polynomial Bayes as the classification algorithm and classification performance as classifier weights. The KNN method yields 0.0375, which equals 3.2452. The polynomial Bayesian method yields 0.9625, which equals 2.5123.

Therefore, it can be seen that the weight of the KNN classifier is 3.2452, and the weight of the polynomial Bayes classifier is 2.5123. The number of erroneous samples of the KNN method is 12,

and the number of erroneous samples of the polynomial Bayes method is 10. When the test sample points are input into the categorized classification, the number of erroneous test samples of the combination method is obtained according to the voting principle. 12, which is the error rate of the KNN method. This method only uses the classification performance of the entire classifier as the weight of the classifier, ignoring the influence of the class on the classifier. Therefore, this paper proposes a category weighted classifier weight calculation method.

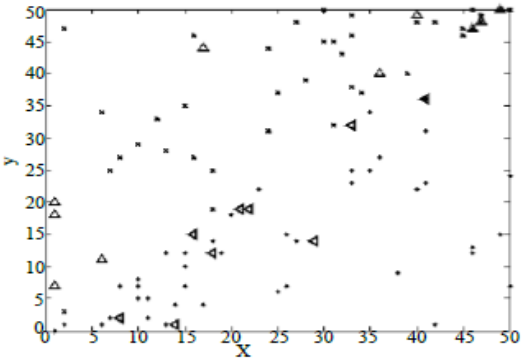

Figure 3. Binary Classification Sample Diagram

The category-weighted classifier considers the influence of category information on the classifier weights. In the above combined classifier, KNN classification performance includes one negative class sample and two positive class samples. Polynomial Bayesian classification performance includes six negative samples. Through the above analysis, the polynomial Bayes classifier has a good recognition rate for positive samples, so we want to increase the weight of polynomial Bayesian positive samples and reduce the classifier weight of negative samples. Therefore, this paper gives different classifier weights with different sample categories. The class-weighted classifier formula is shown in Equation (5). By formula (5), it can be calculated that the positive class of KNN is 0.05, the negative class is 0.025, the positive class is 2.9444, and the negative class is 3.6636. The positive class of polynomial Bayes is 0, the ε of the negative class is 0.15, the positive class is 10 (infinity), and the negative class is 1.7346. KNN and polynomial Bayesian combination method when the test sample is positive, the polynomial Bayes plays a very good role, when the test sample is a negative sample, the KNN method has a greater impact on the combined classifier, according to the voting principle It is concluded that the number of erroneous samples of this combination method is 9. Compared with the previous classifier weight calculation method, the classification effect has been improved.

$$w_{li} = \begin{cases} \alpha, x_i \notin L_l \\ 0, x_i \notin L_l \end{cases} \tag{5}$$

$$\varepsilon = \frac{errorlNum}{textNum} \tag{6}$$

It denotes a test sample that represents the classifier weight of the test sample under category l. errorlNum is the classification error rate under category l is shown in equation (4).

Therefore, through the analysis of the above data, it is concluded that the class weighted classifier weighting method can better represent the classifier weights. The class weighted classifier weighting method was integrated into the multi-class classifier method model, and an improved multi-class classifier model was obtained. The model is shown in Figure 4.

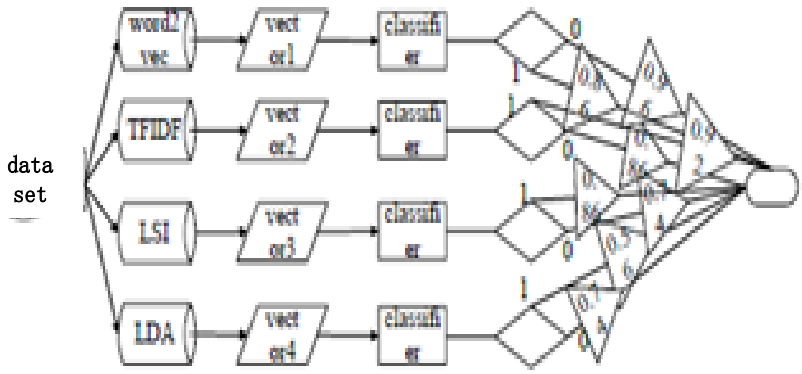

Figure 4. Multi-type binary classifier fusion diagram

Algorithm Steps for Multiple Types of Classifiers

Input: sample training set x_train, sample test set x_test, sample training set label y_train, sample test set label y_test, classifier number classNum.Output: Predicted result matrix.

Calculate word2vec for x_train as class1
Calculate tfidf for x_train as class2
Calculate lsi for x_train as class3
Calculate lda for x_train as class4
Train classifier according to x_train
For i∈{1,2,…,len(x_train)} do
Calculate errorword2vecNum、errortfidfnum、errorlsiNum、errorldaNum for each class according to y_train
End
For i∈{1,2,…,len(x_test)} do
Calculate wli for errorword2vecNum、errortfidfnum、errorlsiNum、errorldaNum according to equation (4)、(5)and(6)
For i∈{1,2,…,len(x_test)} do
For j∈{1,2,…,classNum} do
s[j] += class[j]*w[j][i]
End
Predicted[i] = maximum number index for s is class
End
Return predicted

Experiments and Results Analysis

The experimental platform is based on the anaconda platform, programming language is python language, 4GB memory and 1TB hard disk computer on the experiment.

Experimental Data

In order to verify the performance of multi-category classification fusion method, this paper uses the scene-reviews corpus of NLTK [9] and the common text classification sogou corpus and 20news corpus to verify the experiment [10]. Among them: Sogou corpus and 20news corpus are the most commonly used text categorization corpus, which can be used to test the different performance of the algorithm.

The 20news data is a relatively balanced data set. Movie_ reviews is a sentiment analysis corpus for movie reviews. It is used for categorization work such as sentiment analysis. The use of movie_ reviews can better validate the algorithm. The distribution of data sets is shown in Table 1.

Table 1 Distribution of data sets

Data set name	Category name	Training set/a	Test set/a
Movie_reviews	pos	1600	400
	neg	1600	400
20news	atheism	640	160
	med	792	198
	crypt	794	198
	graphics	800	200
Sogou Corpus	society	1460	365
	entertainment	1460	365

Experimental Analysis

The experiments in this section were divided into three experiments. The first experiment mainly verified the effectiveness of the algorithm. The second experiment focused on the influence of feature dimensions on the algorithm of this paper. The third experiment verified the weights of category-weighted classifiers. The classification method uses the KNC method in the sklearn library in the python language as the classification method (the k value is 10), the feature extraction methods include the word2vec, LSI, LDA and TFIDF methods, filter out feature words less than 30, and use the sample recognition rate as the classification. Evaluation criteria, and the use of 6 fold cross validation method to verify the effectiveness of the algorithm. In this paper, we experimented with multi-type fusion methods and other sub-hybrid methods to verify the performance of our algorithm. The specific results are shown in Table 2 below

Table 2 Test data

	Movie_reviews(%)	20news(%)	100	300	500	Classification performance	Category weighting
LDA	57.58	92.06	57.75	57.58	57.50	95.11	96.21
TFIDF	70.13	93.25	68.70	70.13	68.75	95.07	95.57
LSI	69.46	93.85	68.95	69.46	68.75	95.80	95.98
word2vec	62.71	79.89	61.65	62.71	61.55	89.59	91.74
LDA+TFIDF	65.33	93.72	64.50	65.33	64.45	89.59	91.74
LSI+TFIDF	71.21	94.44	71.00	71.21	69.85	94.75	95.57
LSI+LDA	65.75	93.32	64.90	65.75	64.30	89.59	91.78
word2vec+TFIDF	69.58	89.15	68.40	69.58	68.30	95.98	95.48
word2vec+LDA	61.54	87.37	60.55	61.54	60.55	89.73	91.78
word2vec+LSI	69.21	88.49	69.35	69.21	68.60	95.80	95.57
LSI+LDA+TFIDF	70.75	95.44	70.20	70.75	69.30	94.75	94.16
word2vec+TFIDF +LDA	69.38	95.11	68.55	69.38	68.00	89.59	95.11
word2vec+LSI +TFIDF	70.75	94.58	70.45	70.75	69.85	96.07	96.94
word2vec+LSI +LDA	69.38	95.30	68.60	68.75	68.05	96.07	96.85
word2vec+LSI +TFIDF+LDA	73.34	96.49	72.15	73.00	70.60	97.40	98.22
average value	67.74	92.16	67.05	67.65	66.56	93.66	94.85
The minimum recognition rate of this algorithm	2.13	1.06	1.15	2.13	0.75	1.32	1.28

Experimental Comparison of Multi-classifier Fusion Methods

This part of the data is in the second and third columns of the above table. The 20news and movie_reviews datasets are used for experiments. 20news is a relatively balanced dataset focusing on multivariate classification, and movie_reviews is a balanced dataset focusing on binary classification and specific scene classification. The feature dimension is a 300-dimensional feature.

As it can be seen from the above table 2, 3 column, the average recognition rate of multivariate classification 20news is 92.16% higher than 67.74% of movie_reviews. This is because movie_reviews is a professional sentiment analysis dataset and 20news categorizes datasets for text, and the fields of application are different. Compared with the multi-category classifier method and the subordinate method of multi-class classifiers, movie_reviews achieves good results, with the lowest increase of 2.13% and the lowest increase of 20news of 1.06%. This is because 20news is a relatively balanced data set, so the classifier's recognition rate tends to be multi-sample.

Influence of Feature Dimensions on Fusion Classifiers

This part of the data for the 4th, 5th, 6th column use movie_reviews to verify the effect of different dimensions (100, 300, 500) on the classification performance. From the data in the table, it can be seen that when the characteristic dimension of the movie_reviews dataset is 300, the performance of the multi-class classifier is the best; with the continuous increase of the feature dimension, the average recognition rate of the classifier decreases. Because of the continuous increase in the number of features, a large number of zeros will appear in the feature vector space of the presentation document, resulting in sparse text feature space and affecting the effect of the classifier.

Comparison of Experimental Results of Weighted Classification Weights

This section is a column of 7 and 8, and the Sogou data set is used to verify the validity of the category-weighted classifier weights. From the reaction data, it can be seen that the category weighted relative classification performance weighting method has a certain effect. The average accuracy of the algorithm in this chapter is 1.19% higher than the classification performance weighting method, and the classifier fusion method is 0.82% higher than the subordinate method. In the separate feature extraction methods, the classification performance has been improved, only in word2vec+TFIDF and word2vec+. The effect of the LSI and LSI+LDA+TFIDF methods has not improved because the data size of the corpus makes word vectors of the word2vec method not very effective in classification, and the recognition rates obtained by LDA, LSI and word2vec are almost the same, making the word2vec+LSI +TFIDF fusion method and word2vec+LSI+TFIDF+LDA fusion method have almost the same effect, which affects the overall classification performance of the fusion classifier.

Conclusion

In this paper, a single classifier and a single feature extraction method are not very scalable, and a text classification method based on multi-class classifiers is proposed. This paper combines four different types of feature extraction methods to form a multi-type text classification method. For the problem that classifier weight does not consider category information, a category-weighted classifier weight calculation method is proposed. The effectiveness of the proposed algorithm was verified by binary classification and multivariate classification experiments. The next step is how to parallelize the method and reduce the computation time of the model.

References

[1] Burdick D, Calimlim M, Flannick J, etal. MAFIA: a maximal frequent itemset algorithm [J]. IEEE Trans on Knowledge an-d a Engineering, 2015,17(11):1490-1504.

[2] Grahne G,Zhu J. High performance mining of maximal frequent itemsets[C]//Proc of the 6th International Workshop on HighPerformance Data Mining.2003.

[3] Shen Gaohui, Liu Peidong, Deng Zhihong. NB-MAFIAA: mining algorithm of longest frequent itemsets based on N-List [J] .Journal of Peking University: natural Science Edition, No.

2016NB-MAFIAA: 199-209.
[4] Lin Chen, Gu Junzhong. Mining algorithm of maximal frequent itemsets based onNodeset [J] .Computer Engineering, 2016,42 / 12: 204-207 / 216.
[5] Deng Z H, Lv S L. Fast mining frequent itemsets using Nodesets[J].Expert Systems with Applications, 2014, 41 (10): 4505-4512.
[6] Deng Z H, Lv S L. PrePost+: an efficient N-lists-based algorithm for mining frequent itemsets via children–parent equivalence pruning[J].Expert Systems with Applications, 2015, 42 (13): 5424-5432.
[7] Dam, Li K, Fournier-Viger P, et al. An efficient algorithm for mining top-rank-k frequent patterns[J]. Applied Intelligence,2016,45(1):9
[8] He Li, Ding Zhaoyun, Jia Yan et al. Candidate Category Search in Large-scale Hierarchical Classification[J]. Chinese Journal of Computers, 2014, 37(1):41-49.
[9] LI Ronglu, WANG Jianhui, CHEN Xiaoyun et al. Chinese text classification using maximum entropy model[J]. Journal of Computer Research and Development, 2015,42(1):94-101.
[10] Huang Wenming, Mo Yang. Chinese Spam SMS Filtering Based on Text Weighted KNN Algorithm [J]. Computer Engineering, 2017, 43(3): 193-199.

About the Author:

Zeng Meilin, Bachelor, Librarian, The main research direction is library information construction
E-mail:1027205415@qq.com