An improved weight adjustment strategy based on DFS feature selection method

Xin Fu [a], Hongguo Wang [b], Zengzhen Shao [c], Qiuxia Du [d]

School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China

[a] fuxinxinxin@163.com, [b] wang666666@163.com, [c] shaozengzhen@163.com, [d]duqiuxia666@163.com

Abstract

TF-IDF is a commonly used weighting strategy for text classification.When evaluating feature words, it only involves the evaluation of the importance of the current document, and ignores the intrinsic relationship between the feature word and the category information. In this paper,TF-pDFS algorithm is proposed. From the point of view of the category, the novel feature selection method DFS factor is added as the point of feature word category evaluation. At the same time, the relationship between the distance factor and the characteristic word is analyzed, and the distance factor is added to evaluate the importance degree of the characteristic word effectively. Experiments show that TF-pDFS algorithm can effectively improve the classification accuracy.

Keywords

Text classification; feature selection; weight adjustment; TF-pDFS; TF-IDF.

1. Introduction

The rapid development of computer technology makes the number of unstructured text exponentially increased. It is particularly important to look for effective strategies to automatically and efficiently classify these unstructured text .

Text classification [1] as a means of efficient classification, gradually into the line of sight of the researchers. The classification process first needs to be dimensioned [2].The purpose is to reduce all the feature words in the training corpus to the specified dimension, and then the weight adjustment strategy [3] is to specify the weight for each text according to the degree of importance of the feature word in the specified text after the feature word is dimensioned.

In the many weight adjustment strategies, TF-IDF algorithm is first proposed and widely used. TF-IDF is a commonly used weighting technique for information retrieval and text classification. By calculating the frequency of the characteristic words in the specified text to reflect the importance of the characteristic words, and the introduction of the reverse file frequency to exclude the occurrence of common words. In the follow-up experiment and practical application, TF-IDF algorithm has been widely used and has high classification accuracy.

However, the TF-IDF considers only the importance of the feature word in a single document, ignoring the concept between categories. Many researchers are increasingly concerned about this issue, and put forward the corresponding improvement measures. Debole and Sebastiani [4] proposed the concept of STW for the first time, that is, supervised term weighting. The TF-CHI, TF-IG and TF-GR are commonly used for supervised learning weights by text classification. Jiaul H. Paik [5] calculated the document frequency TF according to the law of the length of the feature and the short text. Chen [6] proposed a new statistical model for measuring the proportion of feature words in a category and making full use of the fine-grained feature distribution to obtain an improved TF-IGM algorithm.

Wei Feng, Luo Chen [7] fully consider the semantic information of the characteristic words, rather than just at the statistical level. Lu Yonghe [8] proposed a new weight adjustment strategy by

studying the feature weight of vectorization when constructing the vector correction function TW from the point of view of the importance of feature items and the ability of class distinction. Zhang Yufang [9] by modifying the expression of IDF in TFIDF, to increase those in a class in the frequent occurrence of the weight of the entry; Taiwan and other people [10] focus on the characteristics of the word in the category and the distribution In this paper, we introduce the IDF function of the feature word distribution concentration coefficient, and use the dispersion coefficient to weight the TF-IIDF-DIC weight function. Shen [11] proposed that the BOR-TFIDF algorithm re-adjust the distinguishing of each feature word to each category, that is, to correct the weight of each feature word to improve the classification accuracy. Xu Fengya [12] focus on the influence of class, interclass and low frequency and high frequency characteristics on the classification. On this basis, the construction method and feature weight re-algorithm of low frequency and high weight feature set are proposed, and the algorithm is extended to Hierarchical classification system.

The author's unique strategy to solve the TF-IDF deficiencies, but there are still some problems. In this paper, the deficiency of TF-IDF algorithm and the thought of predecessors are taken into account, and the distribution of feature words in multiple categories is taken into account to better reflect the importance of feature words in specific categories. In this paper, TF-pDFS algorithm is proposed. In this paper, a new dimensionality reduction DFS algorithm is proposed to extract some of the factors as the evaluation index of the importance of the characteristic word. At the same time, the algorithm takes into account the position relation of the characteristic word and introduces the position factor. In this paper, we combine two aspects to improve TF-IDF algorithm.

2. Improved Weight Adjustment Strategy

2.1 TF-IDF algorithm dilemm

TF-IDF [13] full term term-inverse document frequency. Is widely used in the search engine information retrieval model. By extracting the specific characteristics of the sentence as a search word, thus completing the information retrieval. The formula is as follows:

$$TF-IDF = TF \times IDF = tf_t \times log(\frac{N}{df_t})$$

tf_t Indicates the frequency of the document for the individual document, t the frequency of the document and the total number of words. $log(\frac{N}{df_k})$ Inverse text frequency, which N represents the total number of documents. df_t represents the number of documents of t, the two of the business logarithm, the meaning of this formula is that for a specific feature word t, the more the number of documents, the smaller the weight of t.The goal leading this is to remove the impact of public words.

TF-IDF uses the frequency of the characteristic word in a single document as an indicator of its importance. At the same time, for some common words, such as "China", "gentleman" and so on, through the inverse document frequency value excluded, and ultimately get a single document the weight of. Used in information retrieval to achieve great results .

However, the text classification is slightly different from the information retrieval, information retrieval in a single document, and the text classification to see the impact of a single class of documents, the former to the document, the latter as a dividing line. Text classification using a supervised learning strategy, each document specifies the category label, but the application of a wide range of TF-IDF algorithm but did not consider these label information, from a scientific point of view, showing a lot of deficiencies, see in Table 1:

Table 1: Characteristic word category relationship

doc	t1	t2	class
D1	t1		C1
D2	t1		C1
D3	t1	t2	C1
D4		t2	C2
D5		t2	C2
D6			C2

There are six documents in the table, belonging to the C1 and C2 categories. Among them, the characteristic word t1 is distributed in the three documents of C1, and t2 is distributed in one of C1 and C2. Take this as an example, calculate the TF-IDF of both:

$$TF-IDF_{t1}=TF-IDF_{t2}$$

The evaluation feature word t1 is the same as the weight value of t2, that is, t1 is the same as t2. However, we can see that t1 is distributed in the entire document of the C1 category, the highest classification accuracy for C1, and the corresponding t2 classification accuracy is slightly reduced. Therefore, the TF-IDF strategy does not take into account the classification of the characteristic word, the two categories of different characteristics of the same words considered defective.

2.2 Algorithm to improve TF-pDFS

To sum up, the TF-IDF algorithm is the lack of factors that do not join the category, resulting in the imbalance of the weight calculation. In this paper, the DFS algorithm is introduced on the basis of the word information of the known characteristic words as the index of the evaluation of the characteristic word and the category. At the same time, the concept of the position model is introduced to increase the classification accuracy.

2.2.1 Distinguishing feature selector

DFS [14] is a global feature selection algorithm that calculates the final degree of evaluation by calculating the importance of the characteristic words in each category. The degree of importance of a feature in DFS is calculated as follows:

$$DFS(t)=\sum_{i=1}^{M}\frac{P(Ci/t)}{P\left(\bar{t}/Ci\right)+P\left(t/\overline{Ci}\right)+1}$$

among them, c_i represent text classification, $p(c_i/t)$ represent word t the proportion of appearances at c_i, $p\left(\bar{t}/c_i\right)$ represent c_i the proportion of appearances $\bar{t}$, $p\left(t/\overline{c_i}\right)$ represent c_i the proportion of appearances t. DFS selects the characteristic words to follow the four basic principles:

Where c_i denotes the text category, $p(c_i/t)$ denotes the proportion of the characteristic word in c_i, $p\left(\bar{t}/c_i\right)$ denotes the proportion of the characteristic word $\bar{t}$ in the c_i category, and $p\left(t/\overline{c_i}\right)$ denotes the proportion of the characteristic word t in the c_i category. The DFS algorithm selects the characteristic word to follow the four basic principles:

1. Feature words appear frequently in a single category and do not appear in other categories, indicating that the feature word has a high classification accuracy.

2. The characteristic words appear frequently in some categories and do not appear in other categories, indicating that the characteristic words have high classification accuracy.

3. Characteristic words appear frequently in all categories, indicating that the feature word does not have classification effect.

4. Feature words rarely appear in a single category, but not in other categories, indicating that the characteristic word without classification accuracy.

2.2.2 TF-DF Salgorithm

From the above formula, the degree of importance of the total score of a particular term t in a particular category is proportional to the proportion of a particular category and inversely proportional to the proportion of the other categories. Thus, we can abstract the degree of correlation between the characteristic word and the individual category, as shown in the following equation:

$$DFS\left(t,C_i\right)=\frac{P\left(Ci/t\right)}{P\left(\bar{t}/Ci\right)+P\left(t/\overline{Ci}\right)+1}$$

The correlation between the word t and the category C_i is proportional to the proportion of t in C_i, and the proportion of it in other categories is inversely proportional to the proportion of other categories without t.

Since the above equation is the degree of importance of the feature word between classes, it is combined with the frequency of the characteristic word, and the inverse text frequency is eliminated to obtain the new weight adjustment strategy TF-DFS algorithm:

$$TF-DFS(\omega_t)=tf(t)\times DFS\left(t,C_i\right)$$

$$=tf(t)\times\frac{P\left(Ci/t\right)}{P\left(\bar{t}/Ci\right)+P\left(t/\overline{Ci}\right)+1}$$

$tf(t)$ represent the frequency of occurrence of a characteristic word in a single document can effectively calculate the degree of importance for this document. The interference factors include some common words, common words, such as "China" and "world". In order to effectively remove such noise words,

The introduction of $DFS\left(t,C_i\right)$, not only can effectively avoid the interference words, but also take into account the differences between the characteristics of the distribution of different types of differences, so as to effectively solve the IDF imbalance.

2.2.3 TF-pDF Salgorithm

Taking into account the general characteristics of Chinese sentences, corpus important information, such as news corpus, generally put at the beginning or end of the sentence. Thus, the importance of the front and back words in the sentence should be slightly larger than the middle part, and the standard quadratic function distribution.

We use the middle of the sentence as the origin to establish the Cartesian coordinate system, the distance between the word and the word as an independent variable, which will be the second function as the characteristic word position factor. Finally, the TF-pDFS algorithm is proposed to add the distance influence factor. The formula is as follows:

$$TF-pDFS(\omega_t)=tf(t)\times Posit(t)\times DFS\left(t,C_i\right)$$

$$=tf(t)\times\frac{1}{a^2}\times\frac{P\left(Ci/t\right)}{P\left(\bar{t}/Ci\right)+P\left(t/\overline{Ci}\right)+1}$$

In the above equation, a represents the distance between words and words. We assume that the minimum distance between words and words is 1, and $\frac{1}{a^2}$ is the weight of the importance of changing the response feature with distance. And finally get the added category and distance factor TF-pDFS improved weight adjustment strategy.

3. Experiment

3.1 Experimental data set

In this paper, three kinds of corpus are experimentally verified. The first category is the public balance data set provided by Sogou website, including 10 categories, each sample has 1000 samples, a total of 10,000 samples, followed by financial, real estate, education, science and technology, social, fashion, Sports, games, entertainment. The second category is the standard document set Reuters document Reuters-21578 unbalanced data set, this data set has 10 categories, each category sample distribution is uneven, or even a huge difference. The third category is the laboratory project application data set: CNKI egret summary information unbalanced data set, this data set contains two categories, the first is the geographical information containing the egrets, and the second is the non-Geographic information.

3.2 Classifier

The artificial neural network [15] is used as a classifier. Artificial neural network (ANN) referred to as neural network, is a kind of math and neural network structure and function of the mathematical and computational model for the function to estimate or approximate.

Modern neural network is a nonlinear statistical data modeling tool, which is divided into input layer, hidden layer, output layer, between input layer and hidden layer and between hidden layer and output layer. There is a weight on the nerve for tuning, the ultimate goal of training is to calculate the optimal weight of all the nerves.

The training process is as follows: First, for each nerve on the initial value of the initial value; according to the signal forward and error function to calculate the size of the error; according to the error to determine whether the need to adjust the weight, if not, then the most The optimal classifier is used to calculate the weight of the output layer and the weight increment of the hidden layer unit according to the obtained value of y. Finally, the optimal neural network classifier is obtained by n iterations.

3.3 Ealuation standard

In order to evaluate the classification effect, this paper selects two kinds of evaluation indexes: MacF1 and MicF1. The MacF1 formula is as follows:

$$MacF1 = \frac{\sum_{k=1}^{C} F_k}{C}$$

$$F_k = \frac{2 \times P_k \times R_k}{P_k + R_k}$$

P_k (precision)the accuracy rate refers to the number of documents that are correctly classified by the number of documents identified by the classifier as the number of documents of that class. γ (Recall) the recall rate refers to the number of documents that are correctly classified divided by the quotient of the total number of documents to be tested. Is a measure of the overall effect of the classification of the commonly used assessment method, the formula shown above.

The MicF1 formula is as follows:

$$MicF1 = \frac{2 \times p \times r}{p + r}$$

ρ precision, γ Recall Indicates the range for all categories.

3.4 Experiment procedure

In the experiment, the experiment was divided into 5 groups according to the category, 4 of them were training set, 1 was the test set, and each experiment was taken as a test set. The cycle test was conducted five times, and the average of all the experiments As the final result of the test.

Experiment set three groups of contrast experiments, MacF1 and MicF1 as a good classification accuracy evaluation index. Since this paper does not deal with the solution of the optimal dimension after dimension reduction, 100,300,500,700,900,1100,1300,1500 is selected as the dimension selection after dimension reduction.

In the Sogou equilibrium data set, the obtained TF-IDF and TF-pDFS contrast macF1 and micF1 as follows:

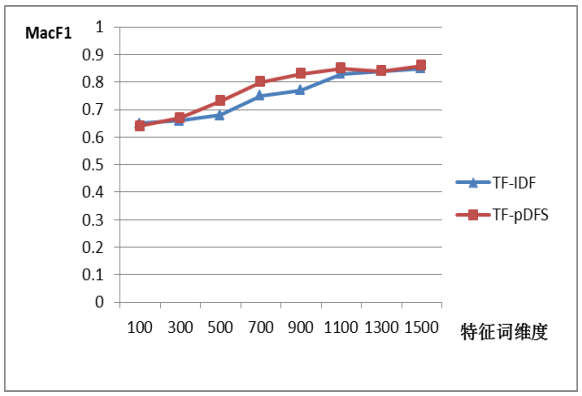

Figure 1 feature word dimension and MacF1 relationship

Figure 1 shows that the classification accuracy of TF-pDFS increases with the increase of TF-IDF, and the effect is gradually increased to 900. After that, the dimension of increasing accuracy is still increasing, But the classification effect of the two tends to coincide. The reason is that different dimensions will affect the weight reduction strategy of weighting strategy, resulting in different classification effects. It can also be seen from the figure that finding the best dimension for different adjustment strategies will be a very interesting direction.

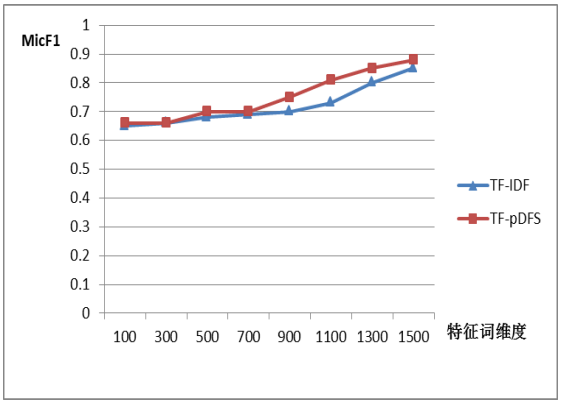

Fig.2 Relationship between feature word dimension and MicF1

Figure 2 can be seen, similar to the MacF1 value in Figure 1, at 1100 when the two reached the maximum difference, and then gradually to the trend of anastomosis.

By comparing TF-IDF with TF-pDFS for MacF1 and MicF1. It is found that the advantage of TF-pDFS algorithm is gradually improved with the increase of dimension of selected feature words.

The reason is that under the balanced data set, the DFS algorithm can be more effective than the importance of the reaction word.

In the Reuters-21578 Uniformed Data Set, the resulting TF-IDF and TF-pDFS are compared to macF1 and micF1 as follows:

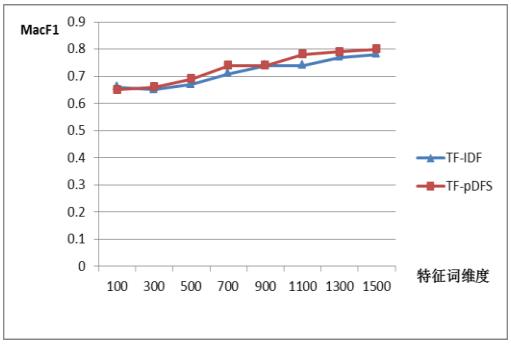

Figure 3 feature word dimension and MacF1 relationship

TF-pDFS and TF-IDF algorithm MacF1 little difference, the former slightly better than the latter. The reason is that the unequal distribution of categories in the document can not fully reflect the relationship between the characteristic word and the category. Therefore, the DFS value is slightly deviated or deviated from the actual target, and the result is shown.

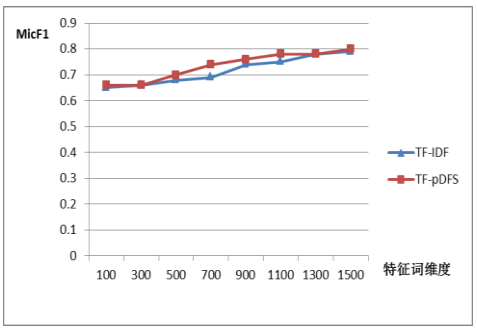

Fig.4 Relationship between feature word dimension and MicF1

Figure 4 can be seen, analogy and Figure 3, the difference between the two MicF1 value is not the same. TF-pDFS is slightly better than TF-IDF only when the feature word dimension increases.

In the CNKI Egret Abstract information unbalanced data set, CNKI Egret Abstract information sample set is a collection of samples collected by the laboratory for the acquisition of bird data. However, CNKI data collected may not be related to the bird data (including the egrets of the word is not necessarily bird data), need to be classified, select one of the bird samples. The sample set we crawled is a binary imbalance sample and the difference is large. So the two algorithms are applied to this sample set for comparison experiments. The obtained macF1 and micF1 values obtained by comparing the obtained TF-IDF with TF-pDFS are as follows

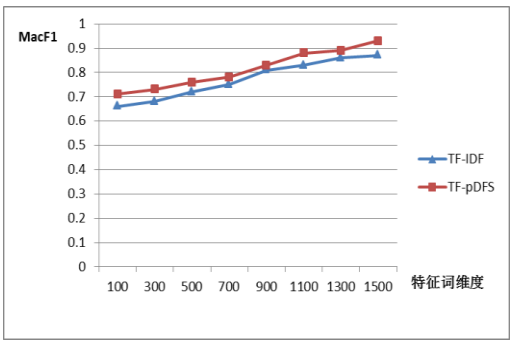

Figure 5 feature word dimension and MacF1 relationship

Figure 5 shows that the TF-pDFS has a higher MacF1 value than the TF-IDF for the binary classification problem, and the accuracy does not change much as the dimension of the feature word increases.

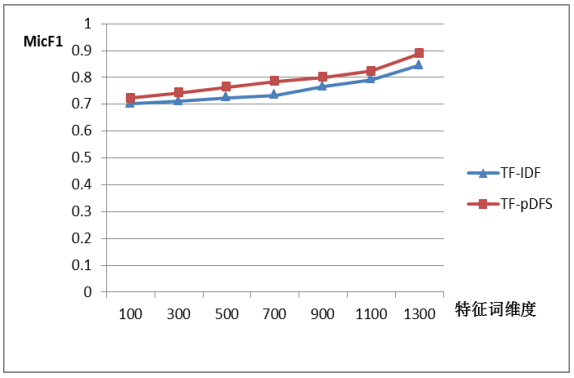

Figure 6 characteristic word dimension and MicF1 relationship

As can be seen from Fig. 6, TF-pDFS has a higher MicF1 value than TF-IDF, and with the increase of the characteristic word dimension, the accuracy does not change much.

4. Conclusion

The shortcoming of TF-IDF is that it only involves the degree of importance of the feature word in the document, while ignoring the influence of the category factor. In this paper, we propose an improved weight adjustment strategy, combined with the feature factor of DFS, and then consider the effect of position factor on classification, and finally get TF-pDFS algorithm.

However, when the algorithm is introduced into the position element, it is only abstract as the general binary function model without deep and detailed reasoning analysis. It is only an optimization strategy, not the best strategy model. At the same time, experiments show that the algorithm performs well on the balanced data set, but the performance is poor under the unbalanced data set. Therefore, the study of the optimal model of the distance factor and the accuracy of the unbalanced data set is the next step must be the primary problem to be solved.

References

[1] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classifycation[C].Advances in neural information processing systems. 2015: 649-657.
[2] Chandrashekar G, Sahin F. A survey on feature selection methods[J]. Computers & Electrical Engineering, 2014, 40(1): 16-28.
[3] Ren F, Sohrab M G. Class-indexing-based term weighting for automatic text classification[J]. Information Sciences, 2013, 236: 109-125.
[4] Debole F, Sebastiani F. Supervised term weighting for automated text categorization[C]. Proceedings of the 2003 ACM symposium on Applied computing. ACM, 2003: 784-788.
[5] Paik J H. A novel TF-IDF weighting scheme for effective ranking[C].Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013: 343-352.
[6] Chen K, Zhang Z, Long J, et al. Turning from TF-IDF to TF-IGM for term weighting in text classification[J].Expert Systems with Applications, 2016, 66: 245-260.
[7] Luo Q, Chen E, Xiong H. A semantic term weighting scheme for text categorization[J]. Expert Systems with Applications, 2011, 38(10): 12708-12716.
[8] LU Yonghe,LI Yanfeng. A Method for Calculating the Weight of Text Feature Item of TF - IDF Algorithm[J].LIBRARY AND INFORMATION SERVICE, 2013, 57(3): 90-95.
[9] ZHANG Yufang,PENG Shiming,LU Jia. Improvement and Application of TFIDF Method Based on Text Classification[J]. COMPUTER ENGINEERING, 2006, 32(19): 76-78.

[10] TAI Deyi,WANG Jun. An Improved Algorithm for Feature Classification of Text Classification [J]. COMPUTER ENGINEERING,2010,36(9): 197-199.
[11] [SHEN Zhibin,BAI Qingyuan. Improvement of Feature Weight Algorithm in Text Classification[J].NANJING NORMAL UNIVERSITY JOURNAL, 2008, 8(4): 95-98.
[12] XU Fengya,LUO Zhensheng. Research on the Improvement of Feature Weights in Text Automatic Classification[J].Computer Engineering and Applications, 2005, 41(1): 181-184.
[13] Yun-tao Z, Ling G, Yong-cheng W. An improved TF-IDF approach for text classification[J]. Journal of Zhejiang University-Science A, 2005, 6(1): 49-55.
[14] Shang W, Huang H, Zhu H, et al. A novel feature selection algorithm for text categorization[J]. Expert Systems with Applications, 2007, 33(1): 1-5.
[15] Fausett L V. Fundamentals of neural networks[M]. Prentice-Hall, 1994.