Post-print of TF-IDF Algorithm Optimization Combining Improved CHI-Square Statistical Method

Authors: Ma Ying, Zhao Hui, Li Wanlong, Pang Hailong, Cui Yan

Date: 2018-05-24T00:00:00+00:00

Abstract

Feature selection and feature weight calculation are two critical components in the text classification process that exert a decisive influence on classification outcomes. To address the limitations of the traditional CHI statistical method, including cases where feature term frequency is negatively correlated with categories and issues concerning the probability of a feature term' s presence in a specific document, this study introduces improvements by incorporating crucial factors such as negative correlation determination and frequency. Furthermore, the TF-IDF algorithm is optimized through integration with semantic similarity computation methods. Experiments conducted on the WEKA platform utilizing KNN (K-nearest neighbor) and Support Vector Machine (SVM) classifiers for microblog sentiment corpus classification demonstrate that the proposed approach yields significant improvements in text classification accuracy.

Full Text

Optimization of TF-IDF Algorithm Combined with Improved CHI Statistical Method

Ma Ying, Zhao Hui†, Li Wanlong, Pang Hailong, Cui Yan (College of Computer Science & Engineering, Changchun University of Technology, Changchun 130012, China)

Abstract: Feature selection and feature weight calculation are two crucial steps in the text classification process that play a key role in determining classification outcomes. To address the limitations of traditional CHI statistical methods—specifically the negative correlation between feature term frequency and category, and the probability issues of feature terms occurring within specific texts—this paper introduces important factors such as negative correlation judgment

and frequency to improve the traditional CHI statistical method. Additionally, the TF-IDF algorithm is optimized by incorporating semantic similarity calculation methods. Experiments conducted on Weibo sentiment corpora using K-nearest neighbor (KNN) and support vector machine (SVM) classifiers in WEKA software demonstrate that the proposed method significantly improves text classification accuracy.

Key words: text categorization; CHI statistics; TF-IDF algorithm; feature selection

---

0 Introduction

With the rapid advancement of the Internet, electronic information has exploded in volume, raising the critical question of how to manage vast amounts of information in a systematic, effective, and organized manner. As a key technology for processing and organizing large-scale data, text classification can largely resolve information disorder, enabling users to quickly obtain valuable information from massive datasets. Consequently, it finds widespread applications in public opinion control, information security, collaborative filtering, product recommendation, and other domains.

Two primary factors influence the final results of text classification: feature selection and feature weight calculation methods. Feature selection involves choosing a subset of valuable terms from a large vocabulary to optimize classification outcomes, while feature weight calculation methods assign weights to feature terms based on statistical analysis, measuring the importance of a particular feature term within a text.

In recent research, Guo Zhengbin optimized the text vector space model through weight and dimensionality adjustments, proposing a novel optimization method for text classification that achieves the goal of vector space optimization. Zhou Qingping combined an improved X2 statistical method with clustering, followed by classification using an improved KNN algorithm, thereby enhancing classification effectiveness. Xu Ming conducted relevant research on feature selection for microblogging, proposing a new chi-square statistical algorithm that demonstrated improved accuracy for microblog information classification when tested with KNN and SVM classifiers.

Among the most commonly used feature selection methods in text classification is the CHI statistical method. However, traditional CHI statistics suffer from two significant shortcomings: first, they exhibit a negative correlation between feature term frequency and category, and tend to select features with relatively low occurrence rates in texts—most of which have weak or no connection to categories; second, they fail to consider the probability of a feature term occurring within a specific text, only accounting for its occurrence across all texts. If a feature term appears concentratedly in most texts of a certain category but

rarely in a few texts of that category, the CHI value may be high, whereas the opposite scenario may yield a low CHI value.

To address these issues, this paper removes cases of negative correlation between feature term frequency and category through positive/negative judgment, introduces important factors such as frequency to improve the traditional CHI statistical method, and incorporates a feature extraction factor . By combining semantic similarity calculation methods with the traditional TF-IDF algorithm, the importance of feature terms within category texts is enhanced, dimensionality is reduced, and ultimately, text classification accuracy is improved.

1 CHI Statistical Method

The CHI statistical method measures the correlation between feature term w and category c using a contingency table .

In the table, A represents the number of texts belonging to category K that contain feature term w; B represents the number of texts not belonging to category K but containing feature term w; C represents the number of texts belonging to category K but not containing feature term w; and D represents the number of texts neither belonging to category K nor containing feature term w. The CHI value is calculated as:

$$\chi^2(w,c) = \frac{N(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)}$$

where $N = A + B + C + D$.

When the CHI value equals 0, it indicates no relationship between feature term w and category c; when the CHI value is larger, it indicates a stronger relationship between them.

Existing research identifies two deficiencies in traditional CHI statistics:

a) Traditional CHI statistical methods exhibit a negative correlation between feature term occurrence frequency and category, and tend to select features with relatively low occurrence proportions in texts. The majority of such features have weak or no connection to categories, with only individual feature words having strong category associations.

b) Traditional CHI statistical methods fail to consider the probability of a feature term occurring within a specific text, only accounting for its occurrence probability across all texts. If a feature term appears concentratedly in most texts of a certain category but rarely appears in a few texts of that category, the CHI value may be high; conversely, the CHI value may be low.

2 Improved CHI Statistical Method

Feature terms and text categories exhibit two types of relationships:

a) When the value of AD - BC is greater than 0 (positive), it indicates a positive correlation between feature term frequency and category. The larger the positive value, the larger its square, and consequently, the larger the CHI value. Therefore, such terms can serve as features for selection.

b) When the value of AD - BC is less than 0 (negative), it indicates a negative correlation between feature term frequency and category. The smaller the negative value, the larger its square, and consequently, the larger the CHI value. Therefore, such terms cannot serve as features for selection.

The traditional CHI statistical formula shows that if the problem of negative correlation between feature term frequency and category remains unresolved, this negative correlation will ultimately affect the CHI value, thereby influencing feature selection results and consequently impacting text classification accuracy. Therefore, this paper addresses this issue by removing cases of negative correlation between feature term frequency and category. The improved formula is:

$$\chi^2(w,c) = \begin{cases} \frac{N(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)} & \text{if } AD-BC>0 \\ 0 & \text{if } AD-BC\le 0 \end{cases}$$

Since traditional CHI statistics do not consider the number of times a feature term occurs within a specific text, only counting its occurrence across all texts, and literature [9] indicates that the key to feature selection lies in the uniform distribution of feature terms within texts of a certain category, this paper introduces frequency , concentration , and dispersion into the traditional CHI formula.

Assuming the training set contains texts $d\_\{j1\},\cdots,d\_\{jk\},\cdots,d\_\{jm\}$ belonging to category $C\_\{j\}$, where $tf\_\{jk\}$ represents the frequency of feature term w in text $d\_\{jk\}$ (1 k m), m represents the total number of texts in a category, $df\_\{j\}$ represents the number of texts in class $C\_\{j\}$ containing feature term w, and n represents the total number of text categories.

a) **Frequency** refers to the proportion of times a feature term appears in texts of a certain category relative to the total number of texts in that category. The frequency of feature term w in category $C\_\{j\}$ is expressed as:

$$\alpha = \frac{\sum_{k=1}^{m} tf_{jk}}{m}$$

b) **Concentration** refers to the proportion of texts containing a particular feature term within a certain category relative to the total number of

texts containing that feature term. The concentration of feature term w in category $C\_\{j\}$ is expressed as:

$$\beta = \frac{df_j}{\sum_{j=1}^{n} df_j}$$

c) **Dispersion** refers to the proportion of texts containing a particular feature term within a certain category relative to the total number of texts in that category. The dispersion is expressed as:

$$\gamma = \frac{df_j}{m}$$

Based on these definitions, the more frequently a feature term w appears concentratedly in most texts of a category, the higher its frequency, concentration, and dispersion, and the greater its contribution to text classification results. Therefore, building upon formula (2), we introduce frequency, concentration, and dispersion to obtain:

$$\chi^2(w,c) = \begin{cases} \frac{N(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)} \times \alpha \times \beta \times \gamma & \text{if } AD - BC > 0 \\ 0 & \text{if } AD - BC \leq 0 \end{cases}$$

3 Traditional TF-IDF Algorithm and Its Improvement

3.1 Traditional TF-IDF Algorithm

TF-IDF is commonly used to measure the importance of a word or term in a corpus. The TF-IDF algorithm was first proposed by Jones [10] and is essentially the product of TF (Term Frequency) and IDF (Inverse Document Frequency).

TF (Term Frequency) measures the frequency of a feature term in a text:

$$TF(w) = \frac{m}{M}$$

where m represents the number of times the feature term appears in text i, and M represents the total number of terms in text i.

IDF (Inverse Document Frequency) measures the discriminative power of a feature term:

$$IDF = \log\left(\frac{N}{n} + 0.01\right)$$

where N is the total number of texts, and n is the total number of texts containing a particular feature term.

The TF-IDF feature extraction function is:

$$W_{ij} = TF_{ij} \times \log\left(\frac{N}{n_{ij}} + 0.01\right)$$

where $W\_\{ij\}$ represents the weight of the feature term, $tf\_\{ij\}$ represents the frequency of a feature term in a particular text, and $n\_\{ij\}$ represents the number of texts containing a particular feature term.

The normalized traditional TF-IDF formula is:

$$W_{ij} = \frac{tf_{ij} \times \log\left(\frac{N}{n_{ij}} + 0.01\right)}{\sqrt{\sum_{j=1}^{M}\left[tf_{ij} \times \log\left(\frac{N}{n_{ij}} + 0.01\right)\right]^2}}$$

3.2 Optimized TF-IDF Algorithm

When using the traditional normalized TF-IDF algorithm to assign weights to feature terms, only the distribution of feature terms within texts is considered, without accounting for the presence of synonyms. The similarity between words is ignored. If this algorithm is used for weight assignment, this characteristic of texts is overlooked. Literature [11] proposes a method for calculating word similarity. Through analysis of knowledge language, we understand that the data structure of knowledge language can be expressed using sets, sememes, and feature structures. Semantic similarity calculation adopts the algorithm from "HowNet" to determine similarity, with the system-set value of being 0.8. This algorithm improves the accuracy of word similarity calculation.

To address the issue of synonyms appearing for feature words in texts, this paper applies semantic similarity calculation to the traditional TF-IDF algorithm, thereby increasing the weight of feature terms in texts and making these feature terms more representative. To this end, we propose a feature extraction factor , which represents the ratio of the number of a particular feature term appearing in a text plus the number of feature terms with similarity greater than to the total number of feature terms. The value of directly affects the importance of feature terms in texts. Its definition formula is:

$$\varepsilon = \frac{a + b}{c}$$

where a represents the number of feature terms $t\_\{ij\}$ existing in text i, b represents the number of feature terms with similarity greater than to feature term $t\_\{ij\}$, and c represents the total number of feature terms.

To improve the accuracy of feature term weights, this paper introduces the feature extraction factor based on semantic similarity calculation to optimize

the traditional normalized TF-IDF algorithm, achieving a combination of form and semantics. The definition formula is:

$$W_{ij} = \frac{tf_{ij} \times \log\left(\frac{N}{n_{ij}} + 0.01\right) \times \varepsilon}{\sqrt{\sum_{j=1}^{M}\left[tf_{ij} \times \log\left(\frac{N}{n_{ij}} + 0.01\right) \times \varepsilon\right]^2}}$$

where $W\_\{ij\}$ represents the weight of the feature term, N represents the total number of category texts, and $n\_\{ij\}$ represents the average of the number of texts containing a particular feature term and the number of texts containing feature terms with similarity greater than to that feature term.

4 Experiments

4.1 Experimental Data and Environment

This paper uses microblog text as experimental data. Compared with traditional web text, microblog text is shorter, with strict constraints on length (limited to 140 characters), and exhibits characteristics of randomness, real-time nature, and uncontrollability [12]. The experiment employed 4,000 Sina Weibo corpora for data analysis. The computer system was Windows 7, using Python programming technology and the Weka 3.6 data mining open-source tool for experimental result comparison. Both KNN and SVM classifiers were used for data testing and analysis in the experiments.

4.2 Evaluation Indicators

Assuming in the classification results for categories, X represents the number of texts where a feature term is correctly classified into a particular category, Y represents the number of texts where a feature term is incorrectly classified into a particular category, and Z represents the number of texts where a feature term is missed from classification. The specific formulas are as follows:

Recall:

$$R = \frac{X}{X + Z} \times 100\%$$

Precision:

$$P = \frac{X}{X + Y} \times 100\%$$

F-value:

$$F = \frac{2 \times P \times R}{P + R} \times 100\%$$

4.3 Experiments and Results

Experiment 1: Microblog sentiment was divided into two aspects: positive sentiment and negative sentiment. Under the same feature dimensions, we compared the performance of three methods—the TF-IDF algorithm combined with improved CHI statistics, traditional CHI statistics, and improved CHI statistics —using the KNN classifier. The experimental results are shown in Table 2 .

Table 2 Comparison of Three Methods Under 500-Dimensional KNN Classifier

As shown in the table, under the same feature dimensions with the KNN classifier, and comparing positive and negative sentiment in microblogs across three metrics—recall R, precision P, and F-value F—the TF-IDF algorithm combined with improved CHI statistics demonstrates improvements over both traditional CHI statistics and improved CHI statistics across all three indicators. The average precision of the TF-IDF algorithm combined with improved CHI statistics is 1.3 percentage points higher than that of traditional CHI statistics, indicating that the proposed method improves the accuracy of microblog sentiment classification.

Experiment 2: We compared the performance of three methods—the IF-IDF algorithm combined with improved CHI statistics, traditional CHI statistics, and improved CHI statistics—under different dimensions using the KNN classifier, specifically comparing their precision. The experimental results are shown in Table 3 .

Table 3 Comparison of Three Methods Under Different Dimensions with KNN Classifier

As shown in Figure 1 [Figure 1: see original paper], regardless of the feature dimension used with the KNN classifier, the TF-IDF algorithm combined with improved CHI statistics achieves more notable classification performance in terms of precision P compared to traditional CHI statistics and improved CHI statistics, with particularly significant accuracy improvements at 200 and 600 dimensional features.

Experiment 3: We compared the performance of traditional CHI statistics, improved CHI statistics, and the TF_IDF algorithm combined with improved CHI statistics across dimensions of 200, 400, 600, 800, and 1000 using both KNN and SVM classifiers. The experimental results are shown in Table 4 .

Table 4 Comparison of Three Methods Under Different Dimensions with SVM Classifier

Comparing the results in Figure 2 [Figure 2: see original paper] with those in Figure 1 [Figure 1: see original paper] reveals that under the same dimensions, the TF-IDF algorithm combined with improved CHI statistics achieves better classification performance with the SVM classifier than with the KNN classifier. This result is consistent with other microblog research findings.

4.4 Experimental Result Analysis

This paper primarily investigates the optimization of the TF-IDF algorithm combined with improved CHI statistics. By improving the traditional CHI statistical method and combining it with the traditional TF-IDF algorithm that incorporates semantic similarity, optimization is achieved. The experimental results show:

As shown in Table 1, under 500 dimensions with the KNN classifier, the improved CHI statistical method achieves an average precision 0.6 percentage points higher than traditional CHI statistics, while the TF-IDF algorithm combined with improved CHI statistics proposed in this paper achieves an average precision 1.3 percentage points higher than traditional CHI statistics.

In Table 2, under 400 dimensions, the accuracy improvement of the IF-IDF algorithm combined with improved CHI statistics is relatively small, possibly due to the influence of synonyms for certain words. At 600 dimensions, accuracy improves again.

Comparing the experimental results from Tables 2 and 3 shows that under the same dimensional features, using the TF-IDF algorithm combined with improved CHI statistics with the SVM classifier yields more favorable classification results than with the KNN classifier, indicating that the SVM classifier is more suitable for the proposed method, ultimately achieving the goal of improving microblog sentiment classification accuracy.

5 Conclusion

Through research and experimentation on text classification technology, this paper proposes an optimization of the TF-IDF algorithm combined with improved CHI statistics. First, the traditional CHI statistical method is improved to address the problem of negative correlation between feature term frequency and category, as well as the probability of a feature term occurring within a specific text. This is then combined with the TF-IDF algorithm optimized with semantic similarity calculation, thereby enhancing the importance of feature terms in texts, achieving dimensionality reduction, and ultimately improving text classification accuracy. Experimental result analysis demonstrates that selecting the TF-IDF algorithm combined with improved CHI statistics for classification under the SVM classifier achieves good classification performance and improves classification accuracy.

References

[1] Gao Yan. Research on related technologies of Weibo sentiment analysis [D]. Beijing: North China Electric Power University, 2014.

[2] Wan Y. Research on Internet Public Opinion Mining Technology Based on Semantic Statistics Analysis [D]. Wuhan University of Technology, 2012.

[3] Xu Yan, Li Jintao, Wang Bin et al. A high performance feature selection method based on differentiating category capability [J]. Journal of Software, 2008, (1): 82-89.

[4] You Fengqin, Zhong Fang, Zhou Zhan. Feature selection method in Chinese multi-category sentiment classification model [J]. Computer Applications, 2016, 36 (S2): 242-246.

[5] Wang Jingzhong, Qiu Cuxiang. Focused topic web crawler based on improved TF-IDF algorithm [J]. Computer Applications, 2015, 35 (10): 2901-2904, 2919.

[6] Guo Zhengbin, Zhang Yangsen, Jiang Yuru. A feature vector optimization method for text classification [J]. Application Research of Computers, 2017, 34 (8): 2299-2302, 2348.

[7] Zhou Qingping, Tan Changgeng, Wang Hongjun, et al. Improved KNN text classification algorithm based on clustering [J]. Application Research of Computers, 2016, 33 (11): 3374-3377, 3382.

[8] Xu Ming, Gao Xiang, Xu Zhigang, et al. Microblog feature extraction method based on improved chi-square statistics [J]. Computer Engineering and Applications, 2014, 50 (19): 113-117, 142.

[9] Xiong Zhongyang, Zhang Pengzhao, Zhang Yufang. Study of text categorization feature selection method based on 2 statistics [J]. Computer Applications, 2008, 28 (2): 513-514, 518.

[10] Jones K S. A statistical interpretation of term specificity and its application in retrieval [J]. Journal of Documentation, 1972, 28 (1): 11-21.

[11] Ren Yaopeng, Chen Lichao, Zhang Jianjun, et al. Research on the calculation methods of feature weights combined with semantics [J]. Computer Engineering and Design, 2010, 31 (10): 2381-2383, 2387.

[12] Zhang Jianfeng, Xia Yunqing, Yao Jianmin. Summary of research on microblogging text processing [J]. Journal of Chinese Information Processing, 2012, 26 (4): 21-27, 42.

Note: Figure translations are in progress. See original paper for figures.