Research and Improvement of CHI Feature Selection in Sentiment Analysis

Li Danyang\textsuperscript{1a}, Fan Huimin\textsuperscript{2b}

\textsuperscript{1}School of Computer Science and Engineering Xi'an Technological University 86-15229895010, 710021

\textsuperscript{2}School of Computer Science and Engineering Xi'an Technological University 86-13891982323, 710021

\textsuperscript{a}821563942@qq.com, \textsuperscript{b}492896361@qq.com

ABSTRACT Feature selection is a very important step in sentiment classification based on machine learning method. This paper will focus on the better CHI feature selection method and make improvements to overcome the shortcomings of the traditional method. The experimental results show that the improved IM-CHI feature selection method improves the F1 value of emotion classification by 4.9% in different feature dimensions, 4.1% in different classifiers, and 2.0% in data sets of different fields. It proves the effectiveness of this method in emotion classification.

1. INTRODUCTION

With the rapid development of the Internet, people freely express their opinions and opinions on the Internet, which contain a lot of valuable information. One of the most representative is e-commerce, people in the major e-commerce websites on the merchandise comments to a large extent affect consumer purchasing decisions [1].

Sentiment analysis, also known as opinion mining, refers to mining the emotional tendencies expressed by reviewers through the analysis of text content. The task of text emotion analysis mainly includes Sentiment classification, emotion information extraction and emotion information retrieval and induction [2]. Machine learning is a common method of text Sentiment classification. Sentiment analysis method based on machine learning [6-8] includes data preprocessing, text representation, feature selection, training classifier and prediction, and feature selection plays an extremely important role in it. Features are part of the characteristics of classified objects and important basis for classification. At the beginning, the selected features will reach tens of thousands or even hundreds of thousands of dimensions, which not only makes the computer operation time very long, but also reduces the accuracy. Feature selection is the use of a certain feature selection algorithm from the beginning of the high-dimensional feature selection part of the information-rich features as a classifier classification features, reduce noise, improve classification accuracy. Reasonable feature selection not only reduces the classification time, but also removes redundant information and improves the classification accuracy. Therefore, feature selection is very important for text emotional classification [3]. At present, the commonly used feature selection algorithms are document frequency (DF), information gain (IG), mutual information (MI), chi-square statistic (CHI), expected cross-tropy (ECE). After analyzing and comparing the feature selection methods such as IG, DF, MI and CHI, Yang draws a conclusion that the classification effect of CHI method is relatively good [4]. But the
algorithm does not involve any form of word frequency information, there is a large word frequency defect, which leads to the accuracy of the method is reduced.

Aiming at the shortcomings of traditional CHI, this paper proposes an improved IM-CHI feature selection method in Sentiment classification, which combines the characteristics of sentiment analysis and overcomes the defect of word frequency. Experiments show that the algorithm proposed in this paper is superior to the traditional CHI algorithm in all aspects for sentiment classification.

2. THE TRADITIONAL CHI ALGORITHM

CHI, proposed by British statistician Pearson, is used to test the independence or relevance of class variables. When a feature is selected using the CHI algorithm, the larger the CHI value of the feature word in the class, the greater the representation of the word for the class. First, suppose the feature word \( w \) and category \( c_i \) are independent. If \( N \) denotes the total number of documents in the document set, \( A \) denotes the number of documents containing the feature \( w \) and belonging to category \( c_i \), \( B \) denotes the number of documents containing the feature \( w \) but not the category \( c_i \), \( C \) denotes the number of documents that do not contain the feature \( w \) but belong to the category \( c_i \), and \( D \) denotes the number of documents that do not contain feature \( w \) nor belong to category \( c_i \). Then, the formula of CHI is shown in formula (1).

\[
CHI(w,c_i) = \frac{N \times (AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}
\]  

(1)

this paper study deeply the traditional CHI algorithm based on sentiment classification, and analyze the following disadvantages:

(1) Word frequency defect. The traditional CHI algorithm only considers whether a feature word appears in a document or not, but does not consider the number of occurrences of the word. If a feature word appears frequently only in a small number of documents of a certain kind, the calculated CHI value is relatively small. This makes the algorithm more likely to choose low-frequency words and reduce the accuracy.

(2) It does not reflect the particularity of sentiment classification. For sentiment classification, the words used to express emotions should be preferred when choosing features. If only a small number of emotional words appear, the calculated CHI value is very small, and it is likely to be excluded in feature selection, which leads to the classification accuracy will be extremely inaccurate.

3. IMPROVED IM-CHI ALGORITHM

Aiming at the two shortcomings of traditional CHI algorithm, this paper adds word frequency information calculation and increases the amount of emotional word information calculation, and improves the algorithm successfully. The improved feature selection algorithm is named IM-CHI algorithm.

In order to solve the problem of word frequency defect in traditional algorithm, this paper proposes the computational complexity \( CHI_{\text{freq}} \), which is used for reference to the traditional Chi-Square statistics, as shown in Formula (2).

\[
CHI_{\text{freq}}(w,c_i) = \frac{N \times C'' \times (F' \times (C'-F') \times C')^2}{WF \times (N-WF)}
\]  

(2)

In formula(2), \( w \) is the feature word, \( c_i \) is the category, \( N \) is the number of the feature words contained in all categories, \( C \) is the number of the feature words contained in class \( c_i \), \( C'' \) is the number of the feature words contained in other categories, \( F \) is the frequency of the feature word \( w \) in class \( c_i \), \( F' \) is the frequency of the feature word \( w \) in other categories, and \( WF \) indicates that the word \( w \) is the word frequency in all categories.

For certain datasets, \( N, C, \) and \( C' \) correspond to constants, so from the right half of formula (2), it can be concluded that the more frequent a feature word \( w \) is in a certain category, the greater the value of the molecular part, and the greater the calculated value of \( CHI_{\text{freq}} \).
For the problem that traditional algorithms do not embody the particularity of affective classification problem, this paper proposes the affective parameter \( \lambda \) to increase the computational value of affective words. The value of the emotional parameter \( \lambda \) is shown in Formula (3).

\[
\lambda(w) = \begin{cases} 
1, & w \text{ is an emotional word} \\
0, & w \text{ is not an emotional word}
\end{cases}
\]

By querying the sentiment dictionary, if the characteristic word \( w \) is the emotional word, the value of \( \lambda \) is 1, otherwise, the value of \( \lambda \) is 0. Therefore, the improved IM-CHI algorithm is shown in Formula (4) by combining the computational complexity of \( \text{CHI}_\text{freq} \) and affective parameter \( \lambda \).

\[
\text{IM-CHI}(w, C) = \text{CHI}_\text{freq} \times \lambda(w) \times \text{CHI}_\text{freq}
\]

\( \text{CHI} \) is the traditional Chi square statistic value. The improved IM-CHI algorithm adds the calculation of the frequency of the feature words to the traditional chi-square value, which overcomes the shortcoming of the traditional algorithm which tends to select the low-frequency words, and the affective parameter \( \lambda \) increases the calculation value of the affective words. Formula (4) shows that when a feature word is an affective word, the value of lambda is 1, then the chi-square value of the affective word is more than the chi-square value of the non-affective word. This makes the possibility of preferential choice of affective words in feature selection more likely, and has better effect in affective classification.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the effectiveness of the improved algorithm, this paper compares the algorithm from several different perspectives, such as data set domain, feature dimension and classifier. The experiment uses the weighted average F1 value of accuracy and recall rate as the evaluation index. The classifier obtained by using CHI algorithm and IM-CHI algorithm respectively in the training process will be used in the test process. Therefore, two classification results will be obtained in the test process. The validity of the improved algorithm is verified by comparing the F1 values of the two results.

The data set used in this experiment is a total of 16000 internet shopping reviews datasets including Chinese emotion mining corpus provided by Dr. Tan Songbo of the Chinese Academy of Sciences, covering six areas: hotels, computers, books, mobile phones, milk, water heaters. The positive and negative reviews of the data set were randomly selected 20% and randomly sorted as test sets.

4.1 Comparative Experiments Under Different Characteristic Dimensions

In order to find the appropriate feature dimension and verify the change of classification F1 value of CHI algorithm and IM-CHI algorithm under different feature dimension, SVM [5] with better classification effect is used to carry out the experiment. The experimental results are shown in figure 1.

![Figure 1. Comparison test results of different feature dimensions under SVM](image)

Experiments show that the F1 value of the improved IM-CHI algorithm is higher than that of CHI.
algorithm in different dimensions.

Analysis of figure 1 shows that when the dimensions are small, the F1 values of the two algorithms are relatively low, and with the increase of the dimensions, the F1 values show an upward trend. After calculation, the improved IM-CHI algorithm improves the F1 value by 4.9% on average in different feature dimensions. In addition, it can be seen from the graph that the F1 value is higher when the feature dimension is about 11000, and then the F1 value tends to be stable.

4.2 Comparative Experiments Under Different Classifiers

In order to verify the effectiveness of IM-CHI algorithm under different classifiers, this experiment selects Naive Bayes, Logical Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor classifier, and carries on the contrast experiment under the same environment of the feature dimension 11000. The experimental results are shown in Figure 2.

![Figure 2. Comparison results of CHI and IM-CHI under different classifiers](image)

Experiments show that the improved IM-CHI algorithm has higher F1 value than CHI algorithm even if different classifiers are used.

Analysis of figure 2 shows that Bayesian classifier has the worst classification effect, but the F1 value of IM-CHI algorithm has been greatly improved, reaching 9.6%. It can also be seen that among all the classifiers, the F1 value of SVM is the highest and the classification effect is the best. This result also corresponds to the results of Pang [5].

4.3 Comparative experiments under different data sets

To verify the effectiveness of the improved IM-CHI algorithm in different data domains, the data sets are divided into six domains: hotel, computer, book, mobile phone, milk and water heater. The data sets of six domains are shown in Table 1 below. Data sets in each field are balanced corpus.

| Dataset domain | Hotel | computer | book |
|----------------|-------|----------|------|
| Comment number | 3500  | 3780     | 3712 |
| Dataset domain | mobile phone | milk | water heater |
| Comment number | 2316  | 2012     | 682  |

In the environment of SVM classifier and feature dimension 11000, the experiments are carried out by using CHI algorithm and IM-CHI algorithm respectively. The experimental results are shown in Figure 3.
Figure 3. Comparison results of CHI and IM-CHI under different data sets

Experiments show that the F1 value of the improved IM-CHI algorithm is higher than that of CHI algorithm under different datasets.

Figure 3 shows that the F1 value decreases with the decrease of the training dataset size, so the training dataset must be above a certain size to ensure the reliability of classification. At the same time, it can be concluded that the improved algorithm's F1 value is very stable, compared with the CHI algorithm's F1 value increased by an average of 2.0%.

5. CONCLUSION

(1) In this paper, an improved IM-CHI feature selection algorithm is designed to overcome the shortcomings of traditional CHI algorithm.

(2) Compared with the traditional CHI algorithm, the improved IM-CHI algorithm improves the classification F1 value by 4.9% on average in different feature dimensions; under different classifiers, the improved algorithm improves the classification F1 value by 4.1%; under different data sets, the improved algorithm improves the classification F1 value by 2.0% on average.

(3) Experiments show that the IM-CHI algorithm proposed in this paper is superior to the traditional CHI algorithm, but the adaptability of the algorithm to data sets in different fields is not stable, so the performance of the algorithm needs to be further improved.

REFERENCES

[1] Li Piji, Ma Jun, Zhang Dongmei. 2012. Label extraction and sorting in user reviews. Journal of Chinese Information Processing, 26: 14-19,45.

[2] Zhao Yanyan, Qin Bing, Liu Ting. 2010. Text sentiment analysis. Journal of Software, 21: 1834-1845.

[3] Yan Lei. 2017. Sentiment text classification based on improved CHI feature selection. Transducer and Microsystem Technologies, 36: 47-51.

[4] Yang Y, Pedersen J O. 1997. A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, 412-420.

[5] B. Pang, L. Lee, S. Vaithyanathan. 2002. Thumbs up Sentiment Classification Using Machine Learning Techniques. Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia: ACL, 79-86.

[6] Song Jingjing. 2013. Sentiment analysis tendency analysis of Chinese short text. Chongqing University of Technology.

[7] Xu Libo. 2012. Extraction of the collocation relationship between product evaluation objects and emotional words. Beijing University of Posts and Telecommunications.

[8] Hu Cheng. 2016. Design and implementation of emotion analysis system based on dependency syntax. South China University of Technology.