An Improved CHI Feature Selection Method for Chinese Text Classification

Wei Zheng  
*Corresponding author(E-mail:247103210@qq.com)

Haidong Wang

Dongmin Wu*

College of Science, Hebei North University, Zhangjiakou 075000, Hebei, China

Library of Hebei Institute of Architecture and Civil Engineering, Zhangjiakou 075000, Hebei, China

Dean's Office, Chongqing Normal University, ChongQing, 401331, China

Abstract

We Proposed a kind of feature selection method named ICHI based on improved CHI. Through the classified experiment, the result shows that feature extraction effect of CHI method is better than the traditional CHI’s when them is used to select features in SVM and KNN classification, and the ICHI method can enhance the accuracy in text classification and it is fitted to extract feature.

Key words: Feature Selection, CHI, Text Classification

1. INTRODUCTION

The text classification is to assign automatically a new document into one or more predefined classes based on its contents. With the development of WWW, in recent years, text categorization (TC) has become one of the key techniques for handing and organizing text data, especially for Web page and document. Therefore, it is very necessary and meaningful to study the key technology of text categorization for improving the speed and accuracy of categorization.

A major difficulty of text categorization is the high dimensionality of the original feature space. Feature selection is an import method to reduce the amount of feature in text categorization, and its goals are improving classification effectiveness and computational efficiency. Currently, the feature selection method’s principle of operation is that it will compute and score for each feature word using statistical knowledge, according to sort the feature words, then it select some feature whose score is higher to act final document feature. Some well-known methods are document frequency (DF), information gain (IG), expected cross entropy (ECE), the weight of evidence of text (WET), $\chi^2$ statistic (CHI) and so on (Feng and Zheng, 2011; Yang and Pederson, 1997), and it is highly desirable to reduce the feature space without the loss of classification accuracy. In recent years, a growing number of statistical classification methods and data mining learning techniques have been applied in this field. Feature selection methods were analyzed in literature (Shenand Lu, 2006), the performance of feature extraction were compared though experimental, the experimental results show that the IG and CHI, have a better effect. IG has a large amount of calculation relative to other several methods, and CHI method has become a common feature selection in text categorization model for its low time complexity, easy to understand, convenient application (Kou and Cai, 2007; Zheng and Wang, 2007). Therefore, it has a realistic significance that we study in-depth the advantages and disadvantages of CHI to find an effective improvement of the model to improve its application efficiency.

In this paper we propose a feature selection method based on improved CHI, named ICHI derived from the CHI original definition. The ICHI overcome the shortcomings of CHI, such as CHI method exist no independence hypothesis between words and categories and the high computational complexity drawback. Experiments on Chinese text data collection collected by the Fudan University show the performance of ICHI method.

The rest of this paper is organized as follows. Section 2 describes the feature selection method CHI commonly used. Section 3 studies the shortcoming of CHI and Section 4 gives an improved CHI method named ICHI. Section 5 discusses the classifier using in experiment to compare ICHI with other text feature selection methods, and presents the experiment’s results and analysis. In the last section, we give the conclusion and future work.
2. CHI CHI-SQUARE STATISTIC (CHI)

CHI is used to measure the interdependencies between classes and feature $t$. The chi statistic method measures the lack of independence between the term and the category. If term $t$ and category are independent, then the CHI is 0.

$$CHI(t,c_i) = \frac{[P(t,c_i)P(\overline{t},c_i) - P(t,\overline{c_i})P(\overline{t},c_i)]^2}{P(t)p(\overline{t})P(c_i)p(\overline{c_i})}$$ (1)

If there are $n$ classes, then each term value will have $n$ correlation value, the average value calculation for a category as follows:

$$CHI(t) = \sum_{i=1}^{n} P(c_i)\log CHI(t,c_i)$$ (2)

3. SHORTCOMINGS OF CHI

Chi-square statistic (CHI) concept from the of Contingency table test, it can be used to measure the statistical correlation between the characteristics of $t$ and category $c$, as shown in table 1.

<table>
<thead>
<tr>
<th>$t_i$</th>
<th>$c_j$</th>
<th>$\overline{c}_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td></td>
<td>$B$</td>
</tr>
<tr>
<td>$C$</td>
<td></td>
<td>$D$</td>
</tr>
</tbody>
</table>

When we measure a related degree between the feature $t$ and category $c$ using statistical methods, we should suppose a hypothesis between $t$ and $c$ in accordance with chi-square distribution with first-order degrees of freedom. $A$ is the document numbers that the documents contains $t$ and belongs to the category $c$. $B$ is the document numbers that the document $s$ contains $t$ and don’t belongs to the category $c$. $C$ is the document numbers that the documents don’t contains $t$ and belongs to the category $c$. $D$ is the document numbers that the document $s$ don’t contains $t$ and don’t belongs to the category $c$. $N$ is total numbers of the experimental data. The correlation degree between term $t$ and category $c$ calculated as shown in the formula (3):

$$\chi^2 = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}$$ (3)

CHI methods maintain the following disadvantages on the formula (2) and (3):

1. There are two conditions in the correlation of categories and feature, which one is positive, and the other is negative. As formula (3) description: if $AD-BC>0$, attribute and category are relevant, which feature document may belong to a certain category, and CHI statistic is greater when a document contains the characteristic more likely to belong to a category. Conversely, if $AD-BC<0$, it is negatively between attribute and category that the feature appear in a related document, but the document may not belong to a category which appears frequently in other categories, but the CHI statistic values may be big, which is a shortcoming of the CHI.

2. The experiment show that CHI’defect is like to look for the low-frequency word. In the actual data set, there are a number of low-frequency words and only a minority of these low-frequency words has a strong correlation with the category. Most of the low-frequency words are noise words, which should not be choose into feature subset. A low frequency noise word only appearing in few categories, are negatively related to most categories, but traditional CHI statistical methods often give a high rating, which affect the classification results. When the data set don’t uneven distribution, the droping in it’s Classification accuracy is obvious.

Because of the shortcoming of CHI method, we propose two kinds of improvement measures to construct a modified feature selection method based on CHI.

4. A FEATURE SELECTION METHOD ICHI BASED ON IMPROVED CHI

Though the analysis, we will improved the traditional CHI statistic methods, and remove the negative situation between category and feature. The improved algorithm is as follows:
The CHI values in improving the CHI statistic methods, is the total sum on all categories of statistic. In other words, when the CHI value is calculated, we only need considering the positive situations. For negative correlation situations between the category and feature words, we reset the CHI statistic to 0 and don’t consider it in the summation statistics, in addition. From related research (Luo, 2004; Yang and Liu, 1999), this article summarizes, when we meet the following 3 demands, it is helpful to select features. These requirements are three demands that they are Dispersion Information, Concentration Information and Frequency Information factors.

(1) Concentration Information: when a feature only appear one or few text, not appear many categories, the feature is helpful to text classification and is power to representation category.

(2) Dispersion Information: If a feature in a uniform appears in the category text, it suggest that it has a strong association with the class, that is, a feature appear in a large amounts of texts in a category, the more dispersed, it is more helpful to category.

(3) Frequency Information: The average number that a feature appear in documents of a category is large, and such feature is value for classification.

Concentration, dispersion, frequency calculation formula is as follows:

(1) Concentration Information

\[
CI = \frac{\text{number of documents that category } C \text{ contain feature } t}{\text{the total number of documents of Training set containing feature } t} \tag{5}
\]

(2) Distribution Information

\[
DI = \frac{\text{numbers of documents that category } C \text{ contain feature } t}{\text{numbers of documents of category } C} \tag{6}
\]

(3) Frequency Information

\[
FI = \frac{\text{numbers of words that category } C \text{ contain feature } t}{\text{numbers of documents of category } C} \tag{7}
\]

This paper proposed a new feature selection method named ICHI (Improved CHI) based on the improved CHI, which concentration, dispersion and frequency are introduced into evaluation function, and ICHI assessment function is as follows:

\[
ICH(t, c_i) = \begin{cases} 
P(t, c_i)P(\overline{T}, \overline{c_i}) - P(t, \overline{c_i})P(\overline{T}, c_i) & AD - BC > 0 \\
0 & AD - BC < 0
\end{cases} \tag{4}
\]

5. EXPERIMENT AND ANALYSIS

We use IG, CHI and ICHI method to select useful features, and carry out classification experiment in order to verify the effect of feature selection.

5.1. Data Collections

The experimental data used in this paper is from Chinese natural language processing group in Department of Computing and Information technology in Fudan university (Luo Yong, 2004). The training corpus is “train.rar” which has 20 categories includes about 9804 documents and “test.rar” includes about 9833 documents is used for test. We just choose some of the documents for our experiments because of considering the efficiency of the algorithm. Table 1 shows the specific quantity of samples in each category we chose.

<table>
<thead>
<tr>
<th>Table 2. Experimental Data</th>
<th>Quantity of training documents</th>
<th>Quantity of testing documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1-computer</td>
<td>134</td>
<td>66</td>
</tr>
<tr>
<td>C2-Enviornmet</td>
<td>134</td>
<td>67</td>
</tr>
<tr>
<td>C3-Education</td>
<td>147</td>
<td>73</td>
</tr>
<tr>
<td>C4-Medical</td>
<td>136</td>
<td>68</td>
</tr>
<tr>
<td>C5-Traffic</td>
<td>143</td>
<td>71</td>
</tr>
<tr>
<td>In all</td>
<td>694</td>
<td>345</td>
</tr>
</tbody>
</table>
5.2. Classifier

SVM and KNN classifier were widely application in in text classification experiment in the, and has made good classification effect (Yang and Liu, 1999; Duand Xiao, 2002), which is currently more outstanding of classifier. The SVM classification is achieved in experiment by support vector machine t specific using LIBSVM package (Ma, Li and Teng, 2008), which is develop and design by Taiwan University Lin Zhiren Dr. KNN classifier is achieved through programming.

5.3. Performance Measure

To evaluate the performance of a text classifier, we use F1 measure put forward byrijsbergen(1979)(Yang. 1999). This measure combines recall and precision as follows:

\[
\text{Recall} = \frac{\text{number of correct positive predictions}}{\text{number of positive examples}}
\]  

(9)

\[
\text{Precision} = \frac{\text{number of correct positive predictions}}{\text{number of positive predictions}}
\]  

(10)

\[
F_1 = \frac{2 \times \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})}
\]  

(11)

5.4. Results and Analysis

We have selected feature experment and the feature dimension is the length from the 1000 to 8000. Experimental results have shown in Figure 1, Figure 2, and the best result for each feature selection method as shown in table 2.

![Figure 1](image1.png)

**Figure 1.** The performance of three feature selection methods on SVM

![Figure 2](image2.png)

**Figure 2.** The performance of three feature selection methods on KNN
Table 3. Method of optimum classification results compare

<table>
<thead>
<tr>
<th>classification</th>
<th>Recall</th>
<th>Precision</th>
<th>F1 value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>IG</td>
<td>98.245%</td>
<td>98.249%</td>
</tr>
<tr>
<td></td>
<td>CHI</td>
<td>98.233%</td>
<td>98.289%</td>
</tr>
<tr>
<td></td>
<td>ICHI</td>
<td>99.012%</td>
<td>99.013%</td>
</tr>
<tr>
<td>KNN</td>
<td>IG</td>
<td>97.932%</td>
<td>97.947%</td>
</tr>
<tr>
<td></td>
<td>CHI</td>
<td>98.512%</td>
<td>98.570%</td>
</tr>
<tr>
<td></td>
<td>ICHI</td>
<td>98.550%</td>
<td>98.557%</td>
</tr>
</tbody>
</table>

Through Figure 1, we can see that the F1 values for CHI is greater than IG and CHI whose classification curve is smooth, which result is consistent to literature (Luo Yong, 2004). When the dimension number of feature is 4000, F1 value of classification corresponding IG curve reached 98.247%. The effect of classification is obviously better than CHI and IG, specially the dimension number of feature is 1000 to 3000. When features dimension number 1000 to 3000, and 3000 to 7000, there were no major transitions on ICHI curve node. When the feature dimension equal to 3000 and 7000, the curve turn point and its corresponding category F1 values of more than 98.9%, especially when the feature dimension is 3000, F1 value for 99.012% Figure 1 maximum point. The classification effect corresponding to ICHI is obviously better than the other two feature selection methods. Figure 2 is a classification effect figure corresponds to three selection methods in using KNN classification algorithm. When dimension number of feature take 1000, the classification effect of CHI is better than IG and ICHI methods. When dimension number of feature take 2000 to 7000, F1 curve of ICHI method corresponds to classification is obviously above CHI and IG curve, and F1 value take of maximum value 98.553% on 6000 dimension. The data in table 3 are extreme values of the recall, precision and F1 values using SVM and KNN classifier corresponding to 3 types of feature selection methods. The maximum F1 values for ICHI is greater than IG and CHI.

Figure 1, Figure 2 and table 2 have shown that a new feature selection method ICHI based on improved CHI could extract some category features that is valid in the Chinese text categorization. Using ICHI method can improve the accuracy of classification, and has a good job stability in the feature extraction.

6. CONCLUSIONS

This paper has proposed an improved feature selection method based on CHI, named ICHI. ICHI implemented three principles which are Concentration Information, Distribution Information and Frequency information. The experiment has shows that ICHI is an effective method to extract category characteristics for feature selection, and it can effectively improve the performance of text categorization. In the future, we will continue work on the representation model of 3 principles of and effective combination of influence factors.

Acknowledgements

The work was supported the National Natural Science Foundation of China (11404088), and Natural Science Foundation of Hebei Province (F2015405011).

REFERENCES


Luo Yong(2004)“The research on improved mutual information feature selection method for text categorization”, Fujian computer, 28(4), pp.82-83.
Ma Jianbin,Li Ying, Teng Guifa(1999)“The comparison studies on the algorithm of KNN and SVM for chinese text Classification”, Journal of agricultural University of Hebei, 2008(3), pp.120-123.