An Improved Feature Extraction Algorithm Based on CHI and MI

Guichuan Feng [a], Shubin Cai [b]

School of Computer and Software, Shenzhen University, Shenzhen, Guangdong 518060, China

[a]312342909@qq.com, [b]shubin@szu.edu.cn

Keywords: text classification, feature weight, Chi-square statistic, mutual information.

Abstract. The problem of high-dimensional feature vector in vector space model (VSM) is an important problem in text classification. After preprocessing the initial number of feature items is generally huge, and many of them for text categorization are useless. This will not only increase the running time of the classification algorithm, but also affect the classification of the text. Previous studies after comparison of various feature extraction methods, it is concluded that the conclusion of CHI and MI method is better, but these methods are different degrees of exist some disadvantages. In this paper, we try to analyze the shortcomings of CHI and Ml method, and propose the viewpoint of parameter correction. Finally, the two methods are fused together, so as to achieve the purpose of improving the classification effect.

1. Introduction

Text classification is an important part of text mining. Because of the good data structure ,vector space model (VSM) is widely used for text representation[1]. High dimensional feature vector is an important problem in vector space model (VSM). The general size of the data set has thousands of features, and most of these feature vectors are useless; by reducing the dimension,we can reduce the running time and accuracy of the classification algorithm. Therefore, in recent years, many scholars study on feature selection problem[2]. The common feature selection methods are mainly Document frequency(DF), mutual information (MI), expected cross entropy (ECE), weight of evidence for text (WET), Chi-square statistic (CHI), information gain (IG),etc.

In recent years, many scholars have tended to study on the feature selection problem, Professor Yang of Carnegie Mellon University through the analysis and comparison of DF, MI, IG and CHI and other methods, come to a conclusion that the MI and CHI method have a relatively good result for classification[3] .Moreover, CHI and IG showed good accuracy in many experiments. Some domestic scholars also tend to study on feature selection problem, Zhao Junyang et al proposed a feature selection method based on maximum mutual information maximum correlation entropy; Shang Wenqian et al proposed a feature selection algorithm based on the Gini index.This paper analyzes the advantages and disadvantages of the traditional CHI and MI algorithm, and proposes a feature selection algorithm FCM based on CHI and MI algorithm.

2. Traditional feature selection method

2.1 Chi-square statistic

χ^2 statistic is often used to detect the independence of the two events, and it can be used to measure the correlation between word w_i and category C. The word w_i For the category C of CHI values can be calculated using the following formula:

$$\chi^2(w_i, c) = \frac{N(A_1A_4 - A_2A_3)^2}{(A_1 + A_3)(A_2 + A_4)(A_1 + A_2)(A_3 + A_4)} \quad (1)$$

In Formula (1), N represents the number of documents that are contained in the document, A_1 represents the number of documents that contain word w_i and belongs to the C class, A_2 represents

the number of documents that contain word w_i but does not belong to the C class. A_3 represents the number of documents that does not contain word w_i and belongs to the C class, A_4 represents the number of documents that does not contain word w_i and does not belongs to the C class, The bigger the value χ^2 is, the more relevant to the category C and the word w_i ,when the value of χ^2 is zero, the word w_i and category C are independent of each other.

2.2 Mutual information

Mutual information is a common method for measuring the value of $V(w_j, c) = 0$, The measure of the mutual information is the existence of word w_j or not to determine the amount of information brought about by category C. Assuming there is a word w_j and category C, then the mutual information between w_j and C can be formalized as defined:

$$MI(w_j, c) = \log \frac{p(w_j \cap c)}{p(w_j)\, p(c)} \tag{2}$$

In Formula (2) , $p(w_j \cap c)$ indicates the probability of occurrence of C and w_j simultaneously, p(w_j) indicates the probability of occurrence of C. From the perspective of information theory, the mutual information measure is the amount of information that the word has brought about by category or not. If the distribution of a lexical entry in the category is equivalent to its distribution on all document sets, then $MI(w_j, c)$ =0. When the word is the best feature to determine the category of the category, the maximum value of mutual information.

3. Improvement of feature selection

3.1 The shortcomings of CHI and MI

Traditional Chi and Mi feature selection method only considers the characteristics in the number of all the documents, without considering the number of characters in a document, which means that it only takes into account the characteristics of word document frequency, and do not consider the characteristics of word frequency, also the position of the word, thus overstating the role of low-frequency words[5]. This has resulted in the ability to be weakened, and the most effective characteristics cannot be selected.

3.2 Improved feature selection method based on CHI and Ml

Through the analysis of the shortcomings of the two methods, we propose the following solution: (1) For both methods have the disadvantage, due to neglect the contribution of the frequency of text classification, we introduce a parameter, to improve frequency. We believe that more is to appear in a class for many times, and in other classes appear less items more representative of this category, so the definition of a formula as follows:

$$x_1 = \frac{T_K(t)}{\sum_{k=1}^{m} T_k(t)} \tag{3}$$

Where $T_K(t)$ denotes the number of terms t that appear in the class of k.

(2) For both methods neglect the position of the word, we can take segmentation method to select features of lexical items, the text can be divided into three parts: the title, keywords, sentence and paragraph text body. For a subset of the three parts of the text, CHI and Ml value feature words can have three parts obtained by adding the following formula:

$$MI(w_i) = y_1 MI(w_i)_{(1)} + y_2 MI(w_i)_{(2)} + y_3 MI(w_i)_{(3)} \tag{4}$$

$$\chi^2(w_i) = y_1 \chi^2(w_i)_{(1)} + y_2 \chi^2(w_i)_{(2)} + y_3 \chi^2(w_i)_{(3)} \tag{5}$$

Where $\sum_{i=1}^{3} y_i = 1$, and $y_1 > y_2 > y_3$.

(3) Considering MI methods can not filter the low-frequency words, we can define a threshold,whenthe word frequency is less than this number ,the word would be given up.

(4) On the second drawback of CHI approach, we believe that both characteristics in a class that appears to be more representative of the hook in this category, so the introduction of the concept of intra-class dispersion, it adds a parameter to fix this shortcoming, the parameters can be defined as in the following forms:

$$x_2 = \frac{DT_k(t)}{|D_k|} \tag{6}$$

where $DT_k(t)$ indicate the number of documents of features in class k, $|D_k|$ indicate all the class k number of documents. By improving the above shortcomings, we can define a new feature extraction algroithm FCM Based on CHI and MI:

$$FCM(w_{i,}c_k) = \frac{x_1.x_2.\chi^2(w_{i,}c_k) + x_1.MI(w_{i,}c_k)}{2} \tag{7}$$

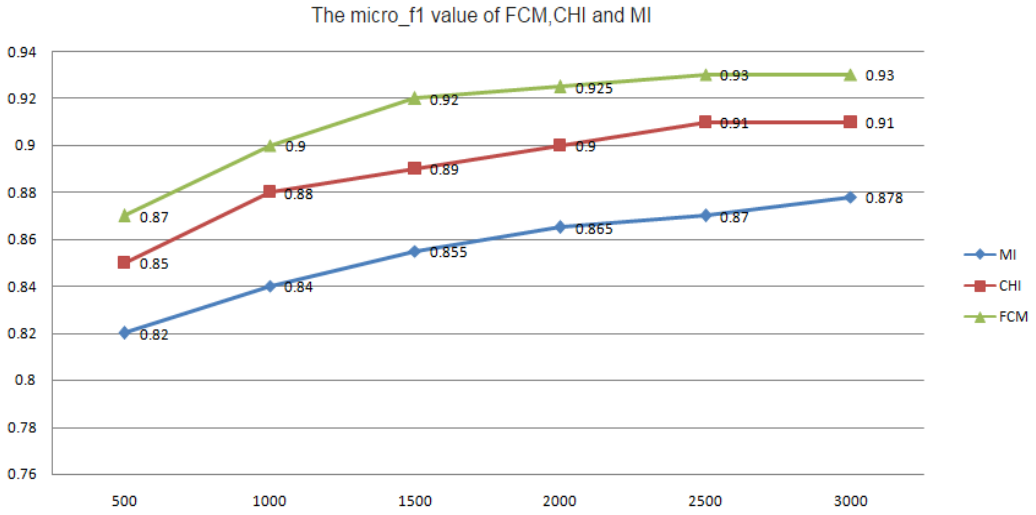
4. Experimental results and analysis

We choose Fudan University Chinese corpus, in order to improve the classification accuracy, after removing the repeat documents, the document is less than 500 words in length or less than 100 in categories, the remaining corpus is as shown in Table 1-1:

Table1-1 Fudan corpus text Distribution

	Environment	Aviation	Agriculture	History	Sport	Computer	Economy	Political	Art
Training set	750	473	816	440	1070	986	1392	756	513
Test Set	741	463	822	447	1066	984	1404	758	508
Total	1491	936	1638	887	2136	1970	2796	1514	1021

We choose KNN as the classification algorithm, the number of features selected from 500-3000, then respectively compare the classification results of FCM, CHI and Mi, Wherein the selected evaluation index is micro_, experiment results are shown in Figure below:



The results can be seen from the chart, the method of FCM can improve text classification results. When the number of feature reaches 2000, text classification evaluation index micro_ tended to a steady value in 2000 at this point, compared with CHI the micro_f1 increased from 90.02% to 92.5%

by using FCM, and 86.5% to 92.5% compared with MI. Compared to the previous method, this method on the classification accuracy has been effectively improved.

5. Summary

By analyzing the shortcomings of the traditional extraction method, we propose a new feature Extraction method. Fudan University Chinese corpus application on KNN classification algorithm for text classification experiments, compared CHI, Ml and FCM, the results show that the proposed method for improving text classification results are valid.

Reference

[1] Liu He. Text Study [D] Classification Problems. Jilin University doctoral dissertation .2009.6.

[2] Hao Xiulan text categorization technology and applied research [D]. Fudan University doctoral dissertation .2008.10.

[3] Zeng yipin have a level research [D] Chinese text sentiment classification. Beijing Jiaotong University, master's thesis .2011.6. [4] Thorsten Joachims, Text Categorization with Suport Vector Machines: Learning with Many Relevant Features, Proceedings of the 10th European Conference on Machine Learning, p. 137-142, April 21-23,1998.

[5] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations. 2010.08, ppl3-16, Beijing, China.

[6] Xipeng Qiu, Feng Ji, Jiayi Zhao and Xuanjing Huang, Joint Segmentation and Tagging with Coupled Sequences Labeling, In COLING, 2012.

[7]JANA N, PETR S, MICHAL H. Conditional mutual information based feature selection for classification task, Proceedings of the 12th International Congress on Pattern Recognition (CIAPR 2007). Berlin: Springer-Verlag, 2007:417-426.