A Feature Selection Based on Relevance and Redundancy

Yonghe Lu*, Wenqiu Liu, Yanfeng Li
Sun Yat-sen University, Guangzhou, China.

* Corresponding author. Tel.: 15360606436; email: luyonghe@mail.sysu.edu.cn

Manuscript submitted January 7, 2015; accepted March 18, 2015.
doi: 10.17706/jcp.10.4.284-291

Abstract: At present, most of the researches on feature selection do not consider the relevance between a term and its own category, the redundancy among terms. In order to solve this problem efficiently, we propose a new feature selection based on analyzing how to measure the relevance and the redundancy, which use Euclidean distance as the similarity calculation method. R2, the new feature selection algorithm, can obtain the optimal feature subset which has considered the correlations between term and category and filtered the redundant terms. Finally, the validity of the new algorithm in feature selection is validated by the classification experiments on Chinese classification corpus by two classifiers, including KNN and Centroid-based classifier.

Key words: Text classification; feature selection; relevance; redundancy.

1. Introduction

Text classification is a key technique that aims at processing large amounts of text data, and it can solve some problems brought by the rapid growth of information. In text classification, the degree of relevance between feature and category and the degree of redundancy among features both have a direct impact on the quality of the classification results [1]. A large number of irrelevant and redundant features can not only influence the classification effect, but also add unnecessary workload.

Feature selection algorithms can quickly reduce the dimensions of feature vector space, simplify the calculation and reduce the training time of models by excluding the irrelevant or redundant features. Common methods of feature selection are document frequency (DF), Chi-square statistic (CHI), information gain (IG) [2] and mutual information (MI) and so on.

To some extent, CHI and IG can effectively select the effective features for text classification [3]. But the methods themselves are insufficient [3]-[6], and do not have a good performance in Chinese text categorization experiments, they need to be modified so as to better suit Chinese text categorization [7]. Therefore, some scholars modified the traditional feature selection methods, and got better results in their experiments. Xiong Zhong-yang, Zhang Peng-zhao and Zhang Yu-fang [6] analyzed the shortcomings of CHI, and applied the frequentness, concentration ratio, degree of dispersion to CHI. Based on MI, Wang Wei-ling, Liu Pei-yu and CHU Jian-chong [8] proposed a feature selection algorithm which consider the correlations both between features and categories and among features. Furthermore, scholars also proposed some modified algorithms based on other areas of knowledge. For example, Shankar, S. and Karypis, G. [9] introduced Gini index in economics, and researched the weighted feature selection with Gini index. Zhenyu Lu *et al.* [10] selected features through counting the words having same semantic. Khan, A., Baharudin, B. and Khan, K. [11] introduced the concept of ontology to extract terms in order to achieve feature selection.

At present, existing feature selection algorithms mainly focus on relevance between feature and category, but ignore the redundancy among features. To overcome this shortcoming, there are some scholars studied the relevance and redundancy. Considering most of the existing feature selections only calculate the relevance, Yu Lei and Liu Huan [12] proposed a new feature selection method combining relevance and redundancy. Inspired by the paper of Yu Lei and Liu Huan [12], Zhou *et al.* [13] proposed joint feature selection algorithm based relevance and redundancy. This method combine document frequency (DF) with IG, MI and CHI respectively, aim at removing redundant features, and retain features that are good for classification, thereby enhance the effect of text sentiment classification. Blum, A. L and Langley P. [14] put forward that the more is the mutual information (MI) between selecting feature and outputting category, the more is the formation used in the classification from features, as a result, the more efficient is classification.

Based on the existing approach, we propose a new feature selection R2 to improve the relevance and reduce the redundancy. First, we calculate the relevance between a word and its category. Second, we calculate the redundancy among words. Last, we combine the relevance and redundancy together.

The difference between joint feature selection method (Zhou Cheng, Ge Bin, Tang Jiuyang, Xiao Weidong. [13]) and R2 is that the former is to joint DF and IG, MI and CHI respectively together and then select features, the latter proposes new methods to calculate the relevance and redundancy.

The paper is organized as follows. In the next section, we propose the method that we research in this paper, and some definitions are given. Section 3 presents the experiments and the experimental results illustrate the efficiency of the algorithm. Finally, we conclude our paper in Section 4.

2. New Feature Selection Algorithm-R2

2.1. Relevant Concepts

In order to formalize the feature selection algorithms based on relevance and redundancy, we define relevance and redundancy as follows.

The relevance of words refers to the degree of relevance between the word and its category, and is used to describe the representative of the word in its category. A higher relevance of word can represents its categories set better than a lower relevance of word, and should has a greater probability of being selected as the feature.

The redundancy of words refers to the degree of relevance between two words. If the relevance between two words is great, that means the two words have a lot of repeat information when deciding the word's category.

In the process of text classification, we usually use the feature vector to describe the text, such as matrix (1).

$$\begin{bmatrix} a_{11},\ldots\ldots, & a_{1i},\ldots\ldots, & a_{1n} \\ a_{21},\ldots\ldots, & a_{2\mathrm{i}},\ldots\ldots, & a_{2n} \\ \ldots\ldots & & \\ a_{j1},\ldots\ldots, & a_{ji},\ldots\ldots, & a_{jn} \\ \ldots\ldots & & \\ a_{m1},\ldots\ldots, & a_{mi},\ldots\ldots, & a_{mn} \end{bmatrix} \tag{1}$$

The matrix (1) means the number of texts is m and the number of features is n in the text set, and each text can be represented by n features. The specific representation of each text is that each row of the matrix consists of an n-dimensional vector, and the value of each dimension a_{ij} is the feature value weighted. Each text can be represented by the feature vectors.

Similarly, we can construct the matrix to represent the words.

$$\begin{bmatrix} \mathrm{b}_{11},\ldots\ldots, & b_{1i},\ldots\ldots, & b_{1n} \\ b_{21},\ldots\ldots, & b_{2\mathrm{i}},\ldots\ldots, & b_{2n} \\ \ldots\ldots & & \\ b_{j1},\ldots\ldots, & b_{ji},\ldots\ldots, & b_{jn} \\ \ldots\ldots & & \\ b_{m1},\ldots\ldots, & b_{mi},\ldots\ldots, & b_{mn} \end{bmatrix} \tag{2}$$

The new matrix (2) has m rows and n columns. The number of training texts in a specific category is n, and m is the word number after segmentation and de-emphasis in training texts in a specific category. Because the text number n of each category in training set is fixed in each classification, we can use n-dimensional vector to describe and compare the words in same category, and that is the row in matrix. Each value b_{ji} in vector represents the frequency number of word j appearing in the text i.

The feature selection in this paper uses the strategy that the text number in a specific category account for the proportion of the total text number in training set, select the appropriate feature number from this category, and the words calculating the relevance and redundancy are in the same category.

According to the discussion above, we can use vectors to describe the words. As shown in equation (3).

$$t_i = \left(b_{i1},\ldots\ldots, b_{ij},\ldots\ldots, b_{in}\right) \tag{3}$$

where t_i represents vector of the i-word in a category. The vector has n dimensions, and the text number in the category is n. b_{ij} represents the frequency number of the i-word appearing in the j-text in the category.

2.2. Calculate the Relevance and Redundancy between Words

The idea of Centroid-based classification is that each category can generate a central feature vector which can approximately represent the vector of the category. Similarly, we assume that there are words $\{t_1, t_2,\ldots\ldots, t_{m-1}, t_m\}$ in the category C, and each word can be represented by the vector in formula (2-3). We can get a virtual word vector by computing all words' center vector in the class, and the vector can be roughly considered as the feature representation of this class. Calculation of the word center vector is shown in the equation (4):

$$C_k = \left(\frac{1}{m}\sum_{i=1}^{m} b_{i1},\ldots\ldots,\frac{1}{m}\sum_{i=1}^{m} b_{ij},\ldots\ldots,\frac{1}{m}\sum_{i=1}^{m} b_{in}\right) \tag{4}$$

where C_k represents the word center vector of class k in training set. This vector has n dimensions, and n is the number of texts in class k. The actual value of each dimension is the corresponding average value of all words in class k.

By calculating the similarity between the word vector and the center vector of all words in class, and we can get a rough relevance of the words, as shown in equation (5):

$$Relevance(t_i, C_k) = \sqrt{\sum_{j=1}^{n}\left(t_{ij} - C_{kj}\right)^2} \tag{5}$$

where $Relevance(t_i, C_k)$ represents the similarity between the vector t_i of word i and the center vector C_k of class *k*. The measure of the similarity is the Euclidean distance, and n is the dimension of the vector (n also is the text number of class k).

In fact, there is distance between the calculated word center vector and the practical center vector of the class, (5) need to be modified. Therefore we proposed the modified equation (6):

$$Relevance(t_i, C_k) = \frac{A}{B+1} \times \sqrt{\sum_{j=1}^{n} \left(t_{ij} - C_{kj}\right)^2} \tag{6}$$

where A represents the text number which contains word i and belongs to class k, B represents the text number which contains word i and not belongs to class k. When the value of A is bigger and the value of *B* is smaller, the relevance between word i and class k is bigger. By adding the weight, the word center vectors can be corrected in some extent.

By now, we can calculate the relevance between words and categories. At next step, we should calculate the redundancy among words in same category.

We calculate the similarity between words within each category at first. We use the Euclidean distance to compute the similarity, and get a similarity matrix (7):

$$\begin{bmatrix} s_{11}, \ s_{12}, \ldots\ldots, \ s_{1j}, \ldots\ldots, \ s_{1m} \\ \ldots\ldots \\ s_{i1}, \ s_{i2}, \ldots\ldots, \ s_{ij}, \ldots\ldots, \ s_{im} \\ \ldots\ldots \\ s_{m1}, \ s_{m2}, \ldots\ldots, \ s_{mj}, \ldots\ldots, \ s_{mm} \end{bmatrix} \tag{7}$$

where m is the total number of word in class k. Since s_{ij} and s_{ji} both describe the similarity between word i and word j, then $s_{ij} = s_{ji}$, (7) is a symmetric matrix actually.

The value of s_{ij} is the Euclidean value between the vector of word i and the vector of word j. The i row in matrix represents the similarity between word i and other words. By summing for each dimension values in i row, then divided by the total number of words m, we get the $Redundancy(t_i)$ of word i. The specific is shown as (8):

$$Redundancy(t_i) = \frac{1}{m} \sum_{j=1}^{m} s_{ij} \tag{8}$$

If the sum of the similarity between a word and other words is smaller, we believe that there is no redundancy between them. Conversely, if the sum of the similarity between words is larger, we believe that there are redundancies between them. When considering whether taking a word as the feature, we should consider the size of the redundancy.

From the discussion above, we can learn that in the process of feature selection, when the relevance is larger and the redundancy is smaller, the word is more likely to be selected as a feature. After considered the relevance and redundancy of the words, we get the feature selection equation (9):

$$R2\left(t_i\right)=\frac{Relevance(t_i,\ C_k)}{Redundancy\left(t_i\right)} \qquad (9)$$

where $R2(t_i)$ represents the value of feature selection function of word i after considering the word relevance and redundancy, C_k represents the word center vector of class k in training set. We will select the features based on the value of $R2(t_i)$, and the word with a large value will be retained, otherwise it will be excluded.

3. Experiments

3.1. Experimental Setting

To verify the validity of feature selection algorithm R2, we compare it with DF, CHI, and IG in the experiments.

The experiment data set is a part of sogou corpus of text classification full version [15]. We select nine categories from the corpus, namely automotive, finance, IT, health, sports, tourism, education, recruitment and military. Each category has 150 texts, and the texts are divided into training text set and testing text set according the ratio of 1:2, namely training text set and testing text set have 450 and 900 texts respectively. The programming language is Java, the programming environment is Eclipse. We use ICTCLAS to do Chinese word segmentation. The number of selected features are 600, 1200, ..., 8400. The feature selection algorithms are DF, IG, CHI and R2. The feature weight calculations are TFIDF and log TFIDF. The classification algorithms are K-nearest neighbor (KNN) and Centroid-based classification [16], and the k in KNN is 10. The evaluation criteria of classification results are Macro-averaging Recall (MR), Macro-averaging Precision (MP) and Macro-averaging F-measure (MF).

3.2. Experimental Results

After using two different feature weight calculations and two different classifiers, we get the macro-averaging F1 of the classification results, and they are shown in Figs. 1-4.

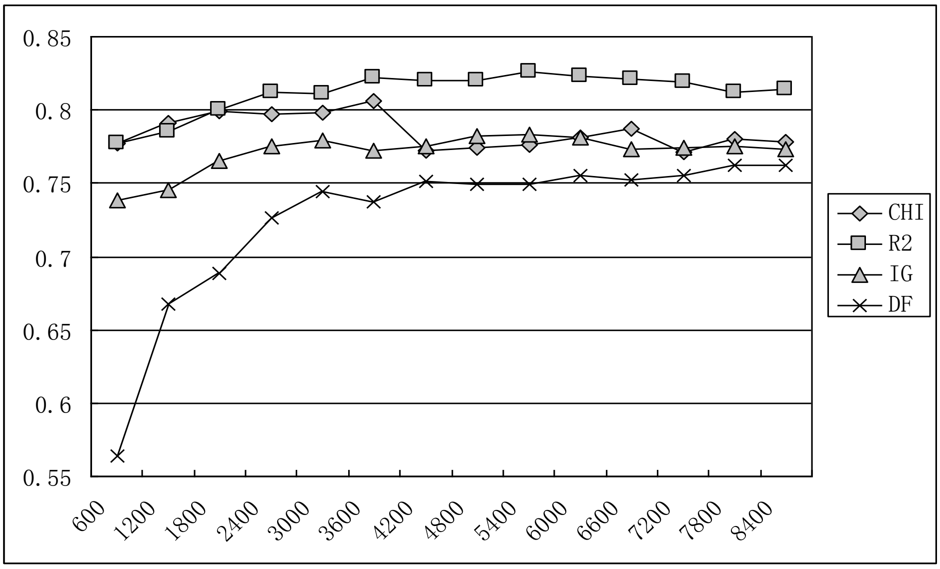

Fig. 1. Comparison of macro-averaging F1--- TFIDF+KNN.

As we can see from the above figures, the feature selection R2 is effective in the classification process. And comparing with the traditional feature selections CHI, IG and DF, the macro-averaging F1 of R2 is better than the others. Especially when the feature number is larger, R2 can make a more comprehensive feature selection.

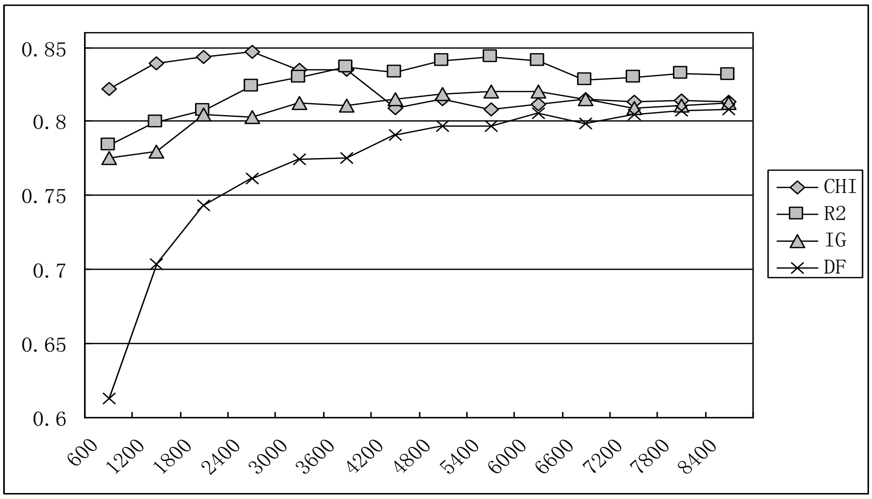

Fig. 2. Comparison of macro-averaging F1--- log TFIDF+KNN.

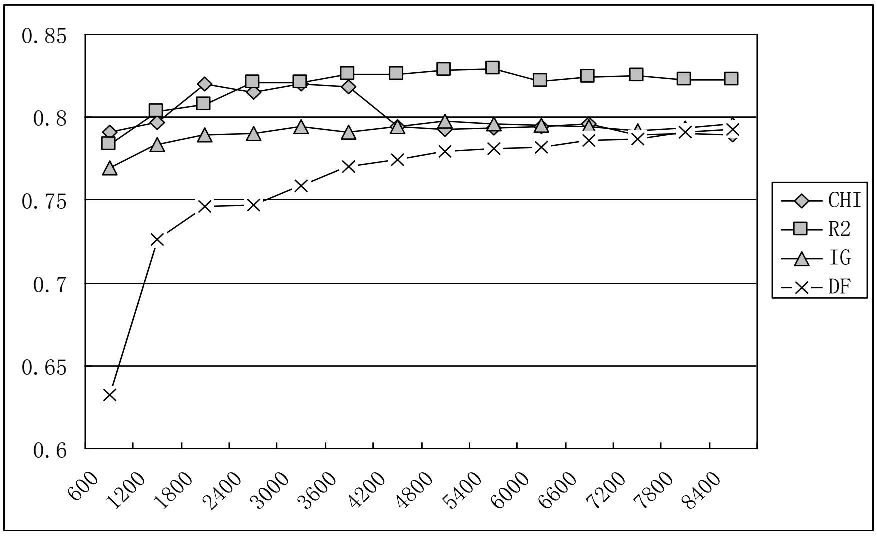

Fig. 3. Comparison of macro-averaging F1--- TFIDF+Centroid-based classification.

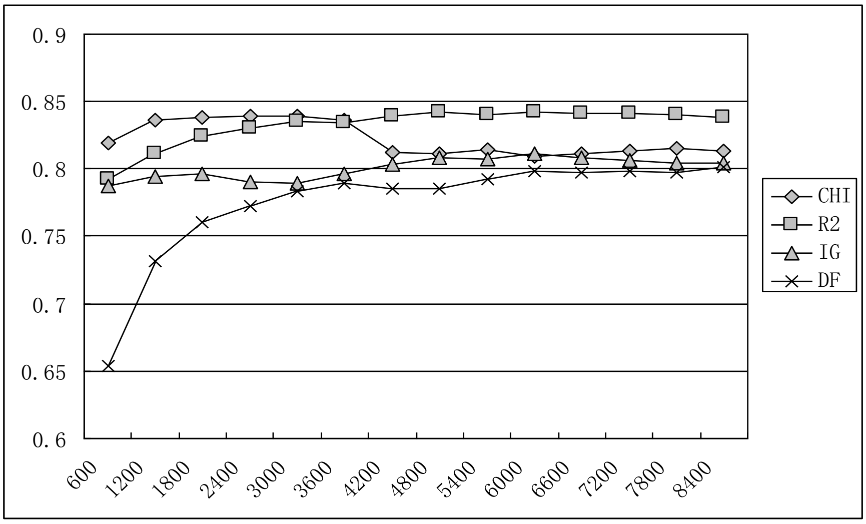

Fig. 4. Comparison of macro-averaging F1--- log TFIDF+Centroid-based classification.

4. Conclusions

In this paper, we focus on the two factors that influence the feature selection---the relevance and redundancy of words. We proposed the feature selection algorithm R2 based on the two factors, and verify the effectiveness of the method with experiments. The major advantage of this algorithm is that it can improve the relevance and reduce the redundancy. Hence, it can improve the classification results.

Acknowledgment

This research is supported by National Natural Science Foundation of China (Grant No. 71373291).

References

[1] Cao, L. (2010). *Feature Selection Research Based on Maximum Relevance Minimum Redundancy*. Unpublished undergraduate dissertation, Yanshan University, Hebei.

[2] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27(1948)*, 379-423, 623-656.

[3] Yang, Y., & Pederson, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the 14th International Conference on Machine learning(ICML' 97). Nashville: Morgan Kaufmann Publishers* (pp. 412-420).

[4] Zhang, F, & Zhang, J-L. (2009). A feature selection method for text information filtering based on statistical frequency. *Library and Information Service*, *53(13),* 116-119.

[5] Liu, Q.-H., & Liang, Z.-Y. (2011). Optimized approach of feature selection based on information gain. *Computer Engineering and Applications*, *47(12),* 130-132.

[6] Xiong, Z.-Y., Zhang, P.-Z., & Zhang, Y.-F. (2008). Improved approach to CHI in feature extraction. *Computer Applications, 28(2),* 513-514.

[7] Dai, L.-L., Huang, H.-Y., & Chen, Z.-X. (2004). A comparative study on feature selection in Chinese text categorization. *Journal of Chinese Information Processing, 18(1),* 26-32.

[8] Wang, W-L., Liu, P.-Y., & Chu, J.-C. (2007). Improved feature selection algorithm with conditional mutual information. *Computer Applications*, *27(2),* 433-435.

[9] Shankar, S., & Karypis, G. (2000). A feature weight adjustment algorithm for document categorization. *Proceedings of The 6th ACM SIGKDD Int'l Conf on knowledge Discovery and Data Mining.* Boston, MA, USA.

[10] Lu, Z., Liu, Y., Zhao, S., & Chen, X. (2000). Study on feature selection and weighting based on synonym merge in text categorization. *Proceedings of IEEE Second International Conference* (pp. 105-109).

[11] Khan, A., Baharudin, B., & Khan, K. (2010). Efficient feature selection and domain relevance term weighting method for document classification. *Proceedings of IEEE Second International Conference on Computer Engineering and Applications* (pp. 398-403).

[12] Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, *5*, 1205-1224.

[13] Zhou, C., Ge, B., Tang, J.-Y., & Xiao, W.-D. (2012). Joint feature selection method based on relevance and redundancy. *Computer Science, 39(4),* 181-184.

[14] Blum, A. L., & Langley, P. (1997) Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97(2),* 245-271.

[15] Sogou Labs. (2006). Text categorization corpus. From http://www.sogou.com/labs/dl/c.html

[16] Liu, P.-L. (2006). *Research on Classification of Chinese Documents Based on Vector Space Model*. Unpublished undergraduate dissertation, Northeast Petroleum University, Daqing.

Yonghe Lu is an associate professor of School of Information Management at Sun Yat-sen University. He is committed to the research on text information analysis and processing, including five directions: text mining, intelligent information processing, custom text classification and clustering, semantic analysis, and public opinion analysis. His research has been published in the Expert Systems with Applications, Applied Mechanics and Materials, New Technology of Library and Information Service, Library and Information Service, Information Studies: Theory & Application, and Journal of Library Science among other outlets.

Wenqiu Liu is a graduate of School of Information Management at Sun Yat-sen University. Her current research interests include text classification and intelligent information processing.

Yanfeng Li is a graduate of School of Information Management at Sun Yat-sen University. His current research interests include text classification, programming and intelligent information processing.