Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 13, no. 7, November 2017, pp. 1159-1164
DOI: 10.23940/ijpe.17.07.p19.11591164

Text Feature Selection based on Feature Dispersion Degree and Feature Concentration Degree

Zhifeng Zhang[a], Yuhua Li[a], Haodong Zhu[b,*]

[a]*School of Software, Zhengzhou University of Light Industry, Zhengzhou, Henan, 450002, P. R. China*
[b]*School of Computer and Communication Engineering, Zhengzhou University of Light Industry Zhengzhou, Henan, 450002, P. R. China*

Abstract

Text feature selection is one of the key steps in text classification, and thus can affect performance of text classification. In this paper, the feature dispersion degree of between-class documents is first put forward to measure the feature dispersion between categories (the greater its value, the larger the influence of the feature has). The feature concentration degree of within-class documents is then proposed to measure feature concentration in the text of a category (the greater its value, the larger the influence of feature has). Subsequently, a text feature selection method is presented, which uses both of the proposed degrees comprehensively to measure the importance of features. Experimental comparison results show that the proposed feature selection method can often get more representative feature subsets and improve performance of text classification.

Keywords: feature selection; text classification; vector space model; text feature vector

(Submitted on July 17, 2017; First Revised on October 7, 2017; Second Revised on October 15, 2017; Accepted on October 17, 2017)

© 2017 Totem Publisher, Inc. All rights reserved.

1. Introduction

With the popularization of computer and the development of modern information technology, text information increases sharply. Automatic text classification as an effective core technology to process a large amount of text information has attracted great attention to academia. Automatic text classification assigns a document to one or more categories according to the document content as well as its properties [8]. At present, vector space model is generally used to represent a document in text classification model [3]. In this model, the training documents are usually composed of a large number of features, and each feature has a certain degree of effect to text classification. However, these features also involve many related and redundant features. The existence of those redundant features not only increases the cost of time and space of the system, but also greatly limits the choice of classification algorithm and reduces the performance of automatic text classification. So, before the implementation of automatic text classification, redundant features have to be eliminated to reduce operation cost and improve the accuracy and the efficiency of text classification [1,5].

In terms of eliminating redundant features, feature selection is a more effective method. It selects some representative features, which have a larger contribution to classification algorithm, to compose feature subset [10,11]. The general process is to evaluate every feature with scores using an evaluation function, rank them according to these scores, and take those features with high scores to form a feature subset to represent relevant texts.

Currently, the feature selection methods mainly contain document frequency, mutual information, cross entropy, information gain, x^2 statistics, and so on. However, these methods only focus on one aspect of the feature in evaluation phase, which makes the selected feature subset not representative [4,9]. So, in this paper, a new feature selection method is proposed. It considers the feature dispersion degree of between-class documents and the feature concentration degree of

* Corresponding author. Tel.: +86-13592697657; fax: +86-0371-86609559.
E-mail address: zhuhaodong80@163.com.

within-class documents so that it can comprehensively evaluate the selected features; the selected feature subset is more representative to improve the performance of text classification.

2. Problem Description

In text classification, text is often represented by a vector space model, which involves some basic definitions as follows [6,7]:

Definition 1: Text Feature. Text is usually composed of some basic language units, such as words, phrases or a set of these basic language units. These basic language units are collectively referred to as text features. Given a text *T*, which can be represented by a set of features, i.e. $T(t_1, t_2, \cdots, t_n)$, where t_k is feature and $1 \le k \le n$.

Definition 2: Feature Weights. For a text $T(t_1, t_2, \cdots, t_n)$ contains *n* features, where t_k is a feature, $1 \le k \le n$. Usually t_k is endowed with a weight to describe its importance, i.e. $T(t_1, w_1; t_2, w_2; \cdots; t_n, w_n)$. When features are determinate, it is often written $T(w_1, w_2, \cdots, w_n)$ briefly.

Definition 3: Vector Space Model (VSM). Given a text $T(t_1, w_1; t_2, w_2; \cdots; t_n, w_n)$, feature t_k $(1 \le k \le n)$ can generally be repeated several times with a certain order, which greatly increases the difficulty of text analysis. In order to simplify the text analysis, we don't consider the order of t_k in the text, and only think that each feature is different from others. So that $t_1, t_2, \cdots, t_n$ can be considered as an n-dimensional coordinate system and $w_1, w_2, \cdots, w_n$ are corresponding to coordinate values. Consequently, $T(w_1, w_2, \cdots, w_n)$ is regarded as a vector of n dimension to represent the text *T*.

Definition 4: Text Feature Vector. In the vector space model, each text is represented by a vector, whose elements are formed by the weights of features and the vector is called the text feature vector. Text feature vector is a feature description of the text, in a sense it can fully represent the document. For example, there is a text *T* of n features, and its feature vector can be formulated as $T(w_1, w_2, \cdots, w_n)$.

Definition 5. Given a set of m texts, where $T_i (i = 1, 2, \cdots, m)$ represents the i^{th} text feature vector. The set possesses *p* categories, and then the average vector for the entire text set is as follows:

$$\bar{C} = (1/m) \sum_{i=1}^{m} T_i \tag{1}$$

The average vector for *k* categories:

$$\bar{C}_k = (1/m_k) \sum_{i=1}^{m_k} T_i \tag{2}$$

Where m_k denotes the text number belong to category *k*. These two average vectors have the relation $\bar{C} = (1/m) \sum_{k=1}^{p} m_k \times \bar{C}_k$.

3. Feature Dispersion Degree and Feature Concentration Degree

The purpose of feature selection is to find a more representative feature subset, and then use the feature subset to represent the original set of texts. In classical linear discriminant analysis, between-class matrix and within-class matrix are two commonly used objective functions, wherein between-class matrix describes the dispersion degree between various categories of documents, and within-class matrix describes the dispersion degree of within-class documents. Generally speaking, the higher dispersion degree between various categories of documents is, the smaller dispersion degree of within-class documents is, the more beneficial it is for classification system. In this case, the task of the feature selection is to find such a feature subset. If the subset is to be used to represent the training set, then the greater the dispersion degree between various categories of documents is and the greater the concentration degree of within-class documents is, the larger the

importance of feature is. So, the greater a feature's contribution to the dispersion degree of between-class documents is, the more representative the feature is. The greater its contribution to the concentration degree of within-class documents is, the more representative the feature is. Based on this idea, in this paper, two kinds of contributions of feature are defined.

Definition 6. Feature Dispersion Degree of Between-Class Documents: It expresses the contribution of a feature to dispersion degree of between-class documents, the greater its value is, the larger the influence of the feature to discriminate the categories of documents is. It can be described by a formula as follows:

$$\text{Dispersion-degree}(i) = (1/p)\sum_{j=1}^{p}(\bar{C}(i) - \bar{C}_j(i))^2 \tag{3}$$

Where p denotes the number of categories, and $\bar{C}(i)$ denotes the average vector weight of the i^{th} feature in the whole text set, $\bar{C}_j(i)$ denotes the average vector weight of the i^{th} feature in the category j.

Definition 7. Feature Concentration Degree of Within-Class Documents: It expresses the contribution of a feature to concentration degree of within-class documents, the greater its value is, the larger the influence of the feature to represent document class is. It can be described by a formula as follows:

$$\text{Concentration-degree}(i) = (1/m)\sum_{j=1}^{p}\sum_{k=1}^{m_j}(\bar{C}_j(i) - T_{jk}(i))^2 \tag{4}$$

Where m denotes the number of total documents, and p denotes the number of categories, $\bar{C}_j(i)$ denotes the average vector weight of the i^{th} feature in the category j. $T_{jk}(i)$ denotes the weight of the i^{th} feature of k^{th} text feature vector in the category j.

On the basis of above two contributions of a feature, we can get the whole importance of a feature, which can be described by:

$$\text{Importance-degree}(i) = \text{Dispersion-degree}(i) \times \text{Concentration-degree}(i) \tag{5}$$

4. Proposed Feature Selection Method

According to formula (5), we can design a new feature selection method, which is described as follows:

Input: Original feature vector set and categories set.
Output: a feature subset Q.
Step1: according to formula (1), we calculate the average vector C of text set.
Step2: according to formula (2), we calculate the average vectors of various categories.
Step3: according to formula (3), the contribution of a feature to the dispersion degree of between-class documents Dispersion-degree(i) is obtained.
Step4: according to formula (4), the contribution of a feature to the concentration degree of within-class documents Concentration-degree(i) is obtained.
Step5: according to formula (5), we get the whole importance Importance-degree(i) of the feature i.
Step6: we select the top P features according to importance scores to form a subset Q and output.

5. Experimental Verification

5.1. Experimental Corpus

In this paper, we select the Chinese text classification corpus of Fudan University as the experimental corpus, which is constructed by the team of natural language processing of computer information and technology department, and can be downloaded from http://www.nlp.org.cn/categories/default.php? cat_id=16. The experimental corpus contains 20 categories, and is divided into the training set and the test set, and each part contains 20 subdirectories. The same categories of documents are stored in a corresponding subdirectory, meanwhile each storage file contains only a document. All documents are uniquely numbered with the file name. After removing some repeated and damaged documents, only 14378

documents remain, wherein the training set contains 8214 articles, and test set contains 6164 articles, without the repeated cross-category documents, i.e. one document only belongs to one category. The distribution of documents of the corpus is uneven. Among them, there are 1369 training documents for the class Economy of the training set, which has the largest number of documents, and 25 training documents for the minimal class. Meanwhile, there are 11 categories whose number of training documents of categories are less than 100, and the training set and the test set are not overlapped.

5.2. Experimental Environment Settings

We employ the Chinese lexical analysis system (ICTCLAS) developed by Institute of Computing Technology of Chinese Academy of Sciences to carry out word segmentation, and select Weka tool as the experiment platform, which is developed by the university of Waikato, New Zealand. Weka tool includes a range of machine learning algorithms in the field of data mining, such as data preprocessing, classification and regression analysis, clustering, association rules, and visualization, and can be downloaded from the following url: http://www.cs.waikato.ac.nz/ml/weka/. We adopt MATLAB 7.0 to implement numerical calculation.

5.3. Classifier and Evaluation Standards

The proposed method is mainly compared with following three feature selection methods [2]: information gain (IG), x^2 statistics (CHI), mutual information (MI). We use KNN classifier (K is set to 20, and cosine distance is adopted to calculate similarity) implement classification experiment. In order to evaluate performance of these four methods with the changing number of features, we select the micro average and the macro average as the performance evaluation standards.

5.4. Experimental Results and Analysis

Micro-average F_1 and Macro-average F_1 are calculated under different numbers of features, and experimental results are shown in Figure 1 and Figure 2, respectively. Note that in Figure 1 and Figure 2, the numbers from 1 to 15 represent the number of features respectively, i.e. 50, 100, 200, 500, 800, 900, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000. The y-axis represents the corresponding average micro F_1 and macro average F_1 in unit %.

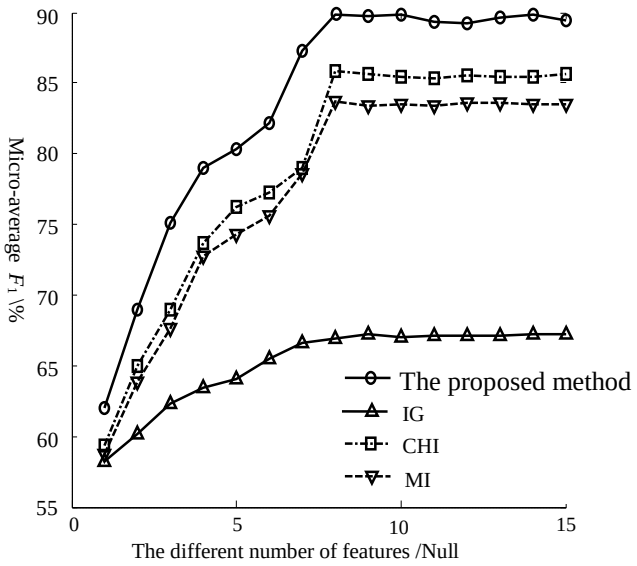

Figure 1. Micro-average F_1 under different number of features

With the changing number of features, the performance change of the classification model can reflect the sensitivity of classification model. As seen from the Figure 1, with the increase of number of features, the micro-average F_1 increases gradually, and achieves a relatively stable level. As seen from the Figure 2, the macro-average F_1 also increases with the growth of number of features but with relatively large fluctuations because the distributions of categories in this selected corpus are extremely uneven. From Figure 1 and Figure 2, it can be seen that KNN classifier has the best performance in top

1500 features for the proposed method; the micro-average F_1 and the macro-average F_1 are about 90% and 80% respectively. KNN classifier for IG has the best performance on selected top 2000 features; the micro-average F_1 and the macro-average F_1 are about 67% and 62%. KNN classifier for CHI has the best performance on selected top 1500 features; the micro-average F_1 and the macro-average F_1 are about 86% and 74%. KNN classifier for MI has the best performance on selected top 1500 features; the micro-average F_1 and the macro-average F_1 are about 84% and 71% respectively. As a result, the overall performances of these four feature selection methods from large to small are the proposed method > CHI > MI > IG. The reason lies in that the proposed method not only considers the influence of the features to the dispersion degree of between-class documents, but also considers the influence of the features to the concentration degree of within-class documents. Therefore, this proposed method is not affected by the distribution of the selected document corpus, and can comprehensively consider the selected features. So, the selected features have better representativeness. At the same time, MI only examines the existing situation of the selected features in selecting feature phase, and CHI considers tow situations of the feature occurring and not occurring, so CHI is superior to MI. Because IG is extremely sensitive to the distribution of the samples, if it is used in the condition of the uneven distribution of samples, the representative of the selected feature set is poorer. In this paper, the distribution of the categories of selected corpus is extremely uneven, so the performance of IG is the worst.

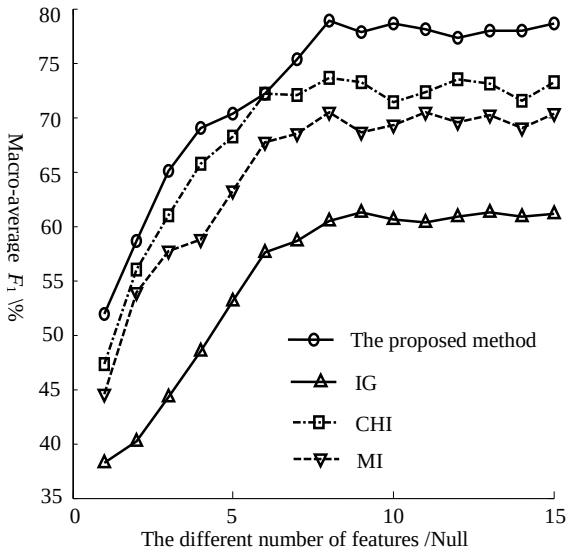

Figure 2. Macro-average F_1 under different number of features

6. Conclusions

This paper defines the feature dispersion degree of between-class documents and the feature concentration degree of within-class documents. Based on them, a new feature selection method is proposed. As demonstrated through experimentations on the corpus of Fudan University and comparisons with three classical methods (MI, CHI and IG), the proposed method has better capability to select the most representative features, and can be used in some knowledge discovery algorithm to reduce time and space complexity. The proposed method not only has application in the text classification, but also provides an idea for other text feature selection methods.

Acknowledgments

The authors would like to thank the editors and the anonymous reviewers for their helpful comments and suggestions, which have improved the presentation. This work is supported in part by the Science and Technology Plan Projects of Henan Province of China under grant No. 152102210357 and No. 152102210149, the Youth Backbone Teachers Funding Planning Project of Colleges and Universities in Henan Province of China under grant No.2014GGJS-084, the Key Science Research Project of Colleges and Universities in Henan Province of China under grant No. 16A520030, the Youth Backbone Teachers Training Targets Funded Project of Zhengzhou University of Light Industry of Henan Province of China under grant No.XGGJS02, the Ph.D. Research Funded Project of Zhengzhou University of Light Industry of Henan Province of China under grant No.2010BSJJ038 and No.2014BSJJ080, and the National Science Foundation of China under grant No.81501548.

References

1. J. Cai, J. Luo, C. Liang, S. Yang, " A Novel Information Theory-Based Ensemble Feature Selection Framework for High-Dimensional Microarray Data", *International Journal of Performability Engineering*, vol. 13, no. 5, pp. 742-753, 2017.
2. A. Destrero, S. Mosci, C. D. Mol,A. Verri, F. Odone, "Feature selection for high-dimensional data", *Computational management science*, vol. 6, no. 1, pp. 25-40, 2009.
3. F. Jiménez, G. Sánchez, J. M. García, et al, "Multi-objective evolutionary feature selection for online sales forecasting", *Neurocomputing*, vol. 234, pp. 75-92, 2017.
4. S. R. Y. Leela, V. Sucharita, B. Debnath, H. J. Kim, "Performance evaluation of feature selection methods on large dimensional databases ", *International Journal of Database Theory and Application*, vol. 9, no. 9, pp. 75-82, 2016.
5. J. H. Liu, Y. J. Lin, M. L. Lin, "Feature selection based on quality of information", *Neurocomputing*, vol. 225, pp. 11-22, 2017.
6. J. N. Meng, H. F. Lin, Y. H. Yu, "A two-stage feature selection method for text categorization", *Computers & Mathematics with Applications*, vol. 62, no. 7, pp. 2793-2800, 2011.
7. M. H. Nguyen, D. F. Torre, "Optimal feature selection for support vector machines", *Pattern Recognition*, vol. 43, no. 3, pp. 584-591, 2010.
8. A. Rehman, K. Javed, H. A. Babri, "Feature selection based on a normalized difference measure for text classification", *Information Processing & Management*, vol. 53, no. 2, pp. 473-489, 2017.
9. T. Sun, S. Y. Qian, H. D. Zhu, "Feature selection method based on category correlation and discernible sets", *Journal of Computational Information Systems*, vol.11, no. 22, pp. 9687-9698, 2014.
10. S. Q. Wang, J. M. Wei, "Feature selection based on measurement of ability to classify subproblems", *Neurocomputing*, vol. 224, pp. 155-165, 2017.
11. H. D. Zhu, H. C. Li, D. Wu, D. S. Huang, B. Wang, "Feature selection method based on feature distinguishability and fractal dimension", *Journal of Information and Computational Science*, vol. 36, no. 5, pp. 6033-6041, 2015.

Biography

Zhifeng Zhang was born in 1978 in Henan Province, China. He received his B.S. degree from Xi'an University of Electronic Science and Technology, Xi'an, Shanxi Province, China, in 2001, and his M.S. degree from Xi'an University of Technology, Xi'an, Shanxi Province, China, in 2006. Since 2006, he has been in the School of Software, Zhengzhou University of Light Industry, Zhengzhou, Henan Province, China, where he is currently an Associate Professor. His major research interests include Cloud Computation, Intelligence Information Processing, and Data Mining.

Yuhua Li is currently a lecturer at School of Software, Zheng Zhou University of Light Industry, Zhengzhou, Henan Province, China. He received his Ph.D. degree in Computer Software and Theory from Sun Yat-sen University, Guangzhou, Guangdong Province, China,in 2014. His current research interests include multimedia information retrieval, machine learning and computational intelligence.

Haodong Zhu was born in Henan Province, China, in 1980. He received the B.S. degree from Lanzhou Jiaotong University, Lanzhou,Gansu Province,China, in 2004, and the M.S. degree from Sichuan University of Science & Engineering, Zigong, Sichuan Province,China,in 2008, and the Ph.D. degree from Graduate University of Chinese Academy of Sciences in 2011. Since 2010, he has been with the faculty of the School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan Province, China, where he is currently an Associate Professor and a master Tutor. His major research interests include Cloud Computation, Intelligence Information Processing, Computing Intelligence and Data Mining.