A Two-stage Feature Selection Method for Text Categorization

Jiana Meng[1,2]
[1]Department of Computer Science and Engineering
Dalian University of Technology
[2]College of Science
Dalian Nationalities University
Dalian, China

Hongfei Lin[1]
[1]Department of Computer Science and Engineering
Dalian University of Technology
Dalian, China

***Abstract*—Feature selection for text classification is a well-studied problem and the goals are improving classification effectiveness, computational efficiency, or both. In this paper, we propose a two-stage feature selection algorithm based on a kind of feature selection method and latent semantic indexing. Traditional word-matching based text categorization system uses vector space model to represent the document. However, it needs a high dimensional space to represent the document, and does not take into account the semantic relationship between terms, which can also lead to poor classification accuracy. Latent semantic indexing can overcome the problems caused by using statistically derived conceptual indices instead of individual words. It constructs a conceptual vector space in which each term or document is represented as a vector in the space. It not only greatly reduces the dimensionality but also discovers the important associative relationship between terms. Because of the too much calculation time of constructing a new semantic space, in this algorithm, firstly we apply a kind of feature selection method to reduce the term dimensions. Secondly, we construct a new reduced semantic space between terms based on latent semantic indexing method. Through some applications involving spam database categorization, we find that our two-stage feature selection method performs better.**

Keywords-feature selection; text categorization; latent semantic indexing; support vector space

I. INTRODUCTION

Feature selection is of considerable importance in pattern classification, data analysis, medical data processing, machine learning, and data mining applications. For many pattern classification problems, a good feature selection method can reduce the cost of feature measurement, and increase classifier efficiency and classification accuracy [1]. Therefore, feature selection can serve as a pre-processing tool of great importance before solving the classification problems.

The most common filters methods include document frequency, information gain (IG), mutual information (MI), Chi-Square statistic (CHI) and odds Ratio. The mentioned above methods are compared in [2]. Odds ratio method is proposed and applied multi-class naïve bayes classifier on reuters-21578 corpus simply in [3]. A new feature selection method by using Support Vector Machine (SVM) is proposed and the different feature selection methods are analyzed in [4]. The method is superior to other methods at specific situations. Two desirable constraints that any reasonable feature selection functions should satisfy are defined and a new feature selection function is presented in [5].

Text representation is an important process to perform text categorization. A major problem of text representation is the high dimensionality of the feature space. The feature space with a large number of terms is not only unsuitable for a classifier but also easily to cause the overfitting problem. The ambiguity meaning of terms can also prohibit the classifier to choose the categories deterministically, and will directly decrease the categorization accuracy. Latent Semantic Indexing (LSI) [6] uses Singular Value Decomposition (SVD) technique to decompose a large term-document matrix into a set of k orthogonal factors, it is an automatic method that can transform the original textual data to a smaller semantic space. Latent Semantic Indexing has been applied to text categorization in many previous works, SVD is used for noise reduction so as to improve the computational efficiency in text categorization in [7], LSA is performed and term-by-document matrix in conjunction with background knowledge in text categorization is expanded in [8], more recently, the supervised LSI [9] has been proposed to improve the performance in text categorization.

This paper proposes a two-stage feature selection algorithm. We combine the feature-based method and semantic-method to reduce the vector space. Firstly, we select features by using a kind of feature selection method such as document frequency, Chi-square and mutual information. The method tends to construct a reductive feature vector space. Secondly, we apply Latent Semantic Indexing to constructs a new conceptual vector space on the basis of the reductive feature vector space. LSI can greatly reduce the dimensionality and discover the important associative relationship between terms. The proposed method not only reduces the number of dimensions drastically, but also overcomes the problems existing in the vector space model used for text representation. We conduct a series of experiments to evaluate the proposed method on two spam filtering corpora. The spam filtering problem can be viewed as a special case of text categorization, with the categories being spam or no-spam. The empirical

978-1-4244-5934-6/10/$26.00 ©2010 IEEE

evaluation results suggest that our proposed method generally outperforms benchmark techniques.

II. Literature Review

A. Support Vector Machine

Support Vector Machine is a new and very popular technique for data classification in the machine learning community. The following concepts are Statistical Learning Theory and structural minimization principle [10]. SVM has been shown to be very effective in the field of text categorization because it can handle high-dimensional data by using kernels. When SVM for pattern classification is used, the basic idea is to find the optimal separating hyperplane that gives the maximum margin between the positive and negative samples. Assume that we have a set of training samples $\mathrm{X}=\{(\mathrm{x}_i, y_i)\}$, where $\mathrm{x}_i \in R^m$ and $y_i \in \{+1,-1\}$ is the corresponding output for the i th training sample (here $+1$ represents spam and -1 stands for legitimate mail). The output of a linear SVM is

$$y = \mathrm{w}\cdot\mathrm{x} - b, \tag{1}$$

where y is the result of classification, w is the normal weight vector corresponding to those in the feature vector x, and b is the bias parameter in the SVM model that is determined by the training process. Maximizing the margin can be achieved through the following optimization problem

$$\text{minimize}\ \frac{1}{2}\|\mathrm{w}\|^2, \tag{2}$$

$$\text{subjected to}\ y_i(\mathrm{w}\cdot\mathrm{x}+b)\geq 1, \forall i.$$

More and more researchers pay attention to SVM-based classifier for spam filtering, since their demonstrated robustness and ability to handle large feature spaces makes them particularly attractive for this work.

III. Two-Stage Feature Selection Method

In the proposed method, feature selection is carried out in two main steps. First, we construct a new reductive feature space by a traditional feature selection method. In the first stage, the original features dimension is decreased from m to t. Second, we select features by LSI method in the basis of the new reductive feature space that constructed in the first stage. In the second stage, the features dimension is decreased from t to k. We combine the feature-based method and semantic-method to reduce the vector space.

A. Construct a New Reductive Feature Space

In order to drop the too much calculation time of constructing a new semantic space from old feature space, we firstly apply a kind of traditional feature selection method to reduce the number of features. The traditional feature selection methods include DF, CHI, MI et al.. We list the methods in table Ⅰ.

A study of the effects of various feature selection methods on an SVM text classifier is described in [11]. An evaluation methodology is proposed for determining the feature selection method or methods that are most likely to provide the best results.

In our method, assumed that $A(m,n)$ is the original term-document matrix. Through the above mentioned feature selection method, we can select $t(t<m)$ features, and so the new term-document matrix is $B(t,n)$.

TABLE I. The Different Feature Selection Method

Method	Formula
DF	$DF(C;t_j)=\sum_{i=1}^{k} df(c_i,t_j)$
CHI	$CHI(C;t_j)=\sum_{i=1}^{k} p(c_i)\frac{N[P(c_i,t_j)\cdot P(\overline{c}_i,t_j)-P(\overline{c}_i,t_j)\cdot P(c_i,\overline{t}_j)]}{\sqrt{P(t_j)\cdot P(\overline{t}_j)\cdot P(c_i)\cdot P(\overline{c}_i)}}$
MI	$MI(C;t_j)=\sum_{i=1}^{k} P(c_i,t_j)\log\frac{P(c_i\mid t_j)}{P(c_i)}$

B. Construct a New Reductive Semantic Feature Space

The semantic feature space method attempts to appropriately capture the underlying conceptual similarity between terms and documents which is helpful for improving the categorization accuracy. This method requires singular value decomposition techniques. Singular value decomposition is a well developed method for extracting the dominant features of large data sets and for reducing the dimensionality of the data. Our corpus can be represented as a term-document matrix $B(t\times n)$ which obtained by constructing the new reductive feature space. Once a term represented by a document matrix is constructed, singular value decomposition is used to decompose it in order to construct a semantic vector space which can be used to represent conceptual term-document associations. The singular value decomposition of B is defined as

$$B = U\Sigma V^T, \tag{3}$$

where $\sum = diag(\sigma_1,\sigma_2,\ldots,\sigma_t)$ is a diagonal matrix of singular values. U and V are the matrices of term vectors and document vectors, respectively. To reduce the dimensionality, we can simply choose the k largest singular values and the corresponding left and right singular vectors. The best approximation of A with a rank-k matrix is given by

$$B_k = U_k\Sigma_k V_k^T, \tag{4}$$

where $\sum_k = diag(\sigma_1,\sigma_2,\ldots,\sigma_k)$ is the first k factors. U_k is comprised of the first k columns of the matrix U and V_k^T is comprised of the first k rows of matrix V^T. The matrix B_k

captures most of the important underlying structure in the association of terms and documents, while ignoring noise due to word choice. SVD is a mathematical concept which is also commonly used in Latent Semantic Indexing methods. LSI is originally proposed as an information retrieval method. In LSI, the effect of this huge dimensional reduction on the data is a muting of the noise caused by synonymy and an enhancing of the latent patterns that indicate semantically similar terms. This means that B_k can actually be a better representation of the data than the original term-document matrix.

When we apply the LSI method, each document in the feature vectors $(1\times t)$ can transform to our desired k - dimensional vectors Therefore, in our experiment, each document is represented by

$$d' = d^T U_k , \tag{5}$$

where d' is the new reduced feature vector and d^T is the feature vector applied by the above mentioned feature selection method. In our experiments all of the training and test examples are represented in this way

IV. Experiments And Results

All experiments that have been done in this paper with software are written in matlab. Standard pre-processing is performed on the raw data. Stop words are eliminated, and stemming is performed with all the data sets. SVMlight [12] package was used as an implementation of SVMs.

A. Spam and Text Classification

In this subsection, we introduce a model for spam categorization. Indeed, electronic mail has gained immense usage in everyday communication for different purposes due to its convenient, economical, fast, and is easy to use nature over traditional methods. However, the volume of spam emails has increased as well. This growing problem on the Internet is costly from different points of view [13]. Several solutions have been proposed to overcome the spam problem. Among the proposed methods, much interest has focused on the machine learning techniques in spam filtering. From the machine learning viewpoint, spam filtering based on the textual content of e-mail can be viewed as a special case of text categorization, with the categories being spam or non-spam.

B. Text Corpora

Many techniques, generally handling the spam problem as a case of text categorization, have been developed to separate spam emails from legitimate ones. In our experiments, we use three large publicly available datasets. The first data set is LingSpam [14], the second data set is collected by Andrew Farrugia [15]. The two data sets represent collect email messages including header, subject, and body text without any alteration, converted to feature vector format indicating the number of occurrences of a given feature (word) in messages. Table Ⅱ shows a description of the three tested data sets.

C. Performance Measures

To evaluate the utility of the various feature selection methods used, we use the F1-measure, a measure that combines precision and recall, two commonly used measures of text categorization performance.

Precision is defined as the ratio of correct classification of documents into categories to the total number of attempted classifications, namely,

$$precision = \frac{true\ positive}{ture\ positive + false\ positive} . \tag{6}$$

Recall is defined as the ratio of correct classifications of documents into categories to the total number of labeled data in the testing set, namely,

$$recall = \frac{true\ positive}{ture\ positive + false\,nagative} . \tag{7}$$

F1-measure is defined as the harmonic mean of precision and recall. Hence, a good classifier is assumed to have a high F1-measure, which indicates that classifier performs well with respect to both precision and recall, namely,

$$F1-measure = \frac{2 * precision * recall}{precision + recall} . \tag{8}$$

TABLE II. Basic Statistics Of The Test Data Sets

Dataset	Ham (%)	Spam (%)	Total
LingSpam	83.4	16.6	2893
Andrew Farrugia	49.75	50.25	5243

D. Results

Figs.1, 2 display the F1-measure performance curves for SVM on LingSpam corpus and Andrew Farrugia corpus after term selection using DF, MI, CHI and LSI, respectively. The x axis is the number of selected features. The y axis is the change in F1-measure value.

As shown in the Figs.1, 2, LSI is superior to DF, MI and CHI feature selection methods slightly. Especially when the selected features number is 50 in Fig.1 and 200 in Fig.2, the F1-measure values of LSI are highest. The best F1-measure values are 0.98 and 0.838. The best performances of DF, MI and CHI in LingSpam are 0.9684, 0.969 and 0.9757, respectively. The best performances of DF, MI and CHI in Andrew Farrugia are 0.8251, 0.8128 and 0.8175, respectively.

Tables Ⅲ, Ⅳ show our two-stage method performances about F1-measure, accuracy and AUC value for the different corpora by using SVM classifier. Moreover, accuracy is used as an indication of overall performance. The AUC value is used

as an evaluation criterion, also. The AUC value is the area under the ROC curve.

Table Ⅲ shows the obtained results about LingSpam corpus when the first step feature selection method is DF method and the size of selection features is 10%. In our experiments in LingSpam corpus, the performances are compared when using the following values for the number of LSI dimensions, k: 25, 50, 75, 100 and 125. For F1-measure and for accuracy, the best performances are equal, 0.9897 are obtained when the number of dimensions is 50. For AUC value, when the number of dimensions is 100, the best performance is obtained. Table Ⅳ shows the obtained results about Andrew Farrugia corpus when the first step feature selection method is DF method and the size of selection features is 5%. In our experiments in Andrew Farrugia corpus, the performances are compared when using the following values for the number of LSI dimensions, k: 25, 50, 75, 100 and 125. For F1-measure and for AUC value, the best performances of 0.8855 and 0.9597 are obtained when the number of dimensions is 100. For accuracy, when the number of dimensions is 125, the best performance of 0.8408 is obtained.

When using the SVM classifier, the F1-measure, accuracy and AUC value for the two-stage feature selection method represent a statistically significant increase from the other feature selection methods, regardless of the text corpus used. In comparison to the other feature selection methods, the two-stage feature selection method appears to perform competitively according to the performances, shows the statistically significant increase consistently in two text categorization tasks.

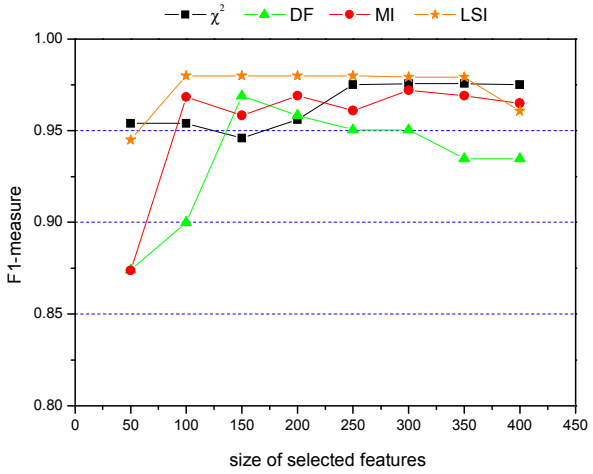

Figure 1. Feature selection and the performance on LingSpam corpus

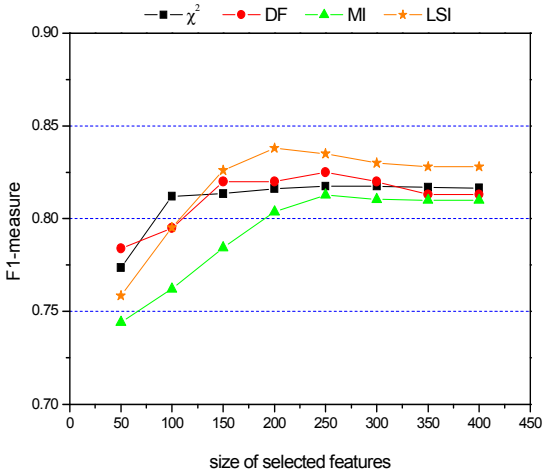

Figure 2. Feature selection and the performance on Andrew Farrugia corpus

TABLE III. PERFORMANCE IN LINGSPAM CORPUS

	k=25	k=50	k=75	k=100	k=125
F1-measure	0.9792	0.9897	0.9362	0.98	0.9691
Accuracy	0.9828	0.9897	0.9794	0.9828	0.9759
AUC	0.9987	0.9997	0.9997	0.9998	0.9991

TABLE IV. PERFORMANCE IN ANDREW FARRUGIA CORPUS

	k=25	k=50	k=75	k=100	k=125
F1-measure	0.8519	0.8573	0.8626	0.8855	0.9691
Accuracy	0.8195	0.8219	0.8354	0.8382	0.8408
AUC	0.9513	0.9572	0.9578	0.9597	0.9578

V. CONCLUSION

The paper proposes a two-stage feature selection algorithm. Firstly, we select features by a kind of traditional feature selection method to reduce the feature numbers observably. Secondly, we apply LSI to construct a new conceptual vector space. The two-stage feature selection method conjugates the vector space model and semantic feature space model. The proposed method not only reduces the number of dimensions drastically, but also overcomes the problems existing in the vector space model used for text representation. Through some applications involving spam database categorization, we find that our two-stage feature selection method outperforms other traditional feature selection methods. Then future works can be devoted to many other problems such as protein sequences in molecular biology, image categorization and text compression.

ACKNOWLEDGMENT

This work is supported by grant from the Natural Science Foundation of China (No.60673039 and 60973068), the

National High Tech Research and Development Plan of China (2006AA01Z151), Doctoral Fund of Ministry of Education of China (20090041110002) and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Jain AK, zongker D., "Feature selection:Evaluation, application, and small sample performance," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol 19, 1997, pp.153-158..

[2] Yang,Y. and Pedersen, J. "A comparative study on feature selection in text categorization,". in 'Proceedings of the 14th International Conference on Machine Learning(ICML'97), Nashville, U.S.A., pp. 412-420, Jul 1997.

[3] Mladenic.D, Grobelnik.M. , "Features selection for unbalanced class distribution and naïve bayes,". In Proceedings of the 16th International Conference on Machine Learning. Bled, Slovenia, pp.258-267. June 1999.

[4] Forman.G. "An extensive empirical study of feature selection metrics for text classification," Journal of Machine Learning Research, vol.3. 2003, pp.1289-1305.

[5] Xu Yan, Li Jin-Tao, Wang Bin, Sun Chun-Ming. "A Category Resolve Power-Based Feature Selection Method," Journal of Software, vol. 19, 2008, pp. 82-89.

[6] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A, "Indexing by latent semantic analysis," Journal of the American Society of Information Science, vol 41, 1990, pp. 391-407.

[7] Y. Yany, "Noise reduction in a statistical approach to text categorization," In Proceeding of the 18th ACM International Conference on Rexorch ond Development in Information Retrieval, New York, 1995, pp. 256–263.

[8] S. Zelikovitz, H. Hirsh, "Using LSI for text classification in the presence of background text," In Proceedings of the tenth international conferencs on Information and knowledge management, ACM Press, 2001, pp. 113-118.

[9] Jian-Tao Sun, Zheng Chen, Hua-Jun Zeng, Yuchang Lu, Chun-Yi Shi, Wei-Ying Ma, "Supervised latent semantic indexing for document categorization," In ICDM, IEEE Press, 2004, pp. 535–538.

[10] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[11] Forman, G., "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research* **3**, 2003,pp. 1289–1305.

[12] http://svmlight.joachims.org/.

[13] .Cranor, L. F., and LaMacchia, B. A. , "Spam! Communications of the ACM," vol 41, 1998, pp.74−83

[14] The Ling-Spam corpus is available from http://www.iit.demokritos.gr/~ionandr/publications.htm.

[15] http://www.csse.monash.edu.au/hons/se-projects/2004/Andrew.Farrugia/.