Accuracy improvement of automatic text classification based on feature transformation and Multi-classifier combination

Xuexian Han[1], Guowei Zu[1,2], Wataru Ohyama[1],
Tetsushi Wakabayashi[1], and Fumitaka Kimura[1]

[1] Faculty of Engineering, Mie University,
1515 Kamihama-cho, Tsu-Shi, Mie, 514-8507, Japan,
http://www.hi.info.mie-u.ac.jp/en/top.html
[2] Toshiba Solutions Corporation, Systems Integration Technology Center
Toshiba Building, 1-1, Shibaura 1-chome, Minato-ku, Tokyo 105-6691, Japan

Abstract. In this paper, we describe a comparative study on techniques of feature transformation and classification to improve the accuracy of automatic text classification. The normalization to the relative word frequency, the principal component analysis (K-L transformation) and the power transformation were applied to the feature vectors, which were classified by the Euclidean distance, the linear discriminant function, the projection distance, the modified projection distance and the SVM. In order to improve the classification accuracy, the multi-classifier combination by majority vote was employed.

1 Introduction

The basic process of automatic text classification is learning a classification scheme from training examples then using it to classify unseen textual documents[1][2]. In this paper, we focus on techniques of feature transformation such as the normalization to the relative word frequency, the principal component analysis and the power transformation to improve the accuracy and the speed of automatic text classification.

1.1 Normalization to relative word frequency

The word frequency is widely used as the basic feature in the statistical text classification approach. Since the absolute frequency depends on the length of the text, the relative frequency:

$$y_i = \frac{x_i}{\sum_{i=1}^{n} x_i} \tag{1}$$

which does not depend on the length is also employed, where x_i is the absolute frequency of word i and n is the number of different words. Because the relative frequency does not depend on the text length, the within-class variance of the relative frequency is smaller than the absolute frequency. Therefore we can expect that separability in the feature space and the classification rate is improved when the relative frequency is employed.

1.2 Power transformation

Another variable transformation, the power transformation [3]:

$$z_i = x_i^v \quad (0 < v < 1) \tag{2}$$

is employed to improve the classification accuracy. This transformation improves the symmetry of the distribution of the frequency $x_i \geq 0$ which is noticeably asymmetric near the origin.

1.3 Dimension reduction by the principal component analysis

Furthermore, it is a critical problem for the statistical classification techniques that the dimensionality of the feature vector can increase together with the lexicon size. To solve the problem we need to employ a statistical feature extraction technique which extracts small number of features with high separability to reduce the feature dimension without sacrificing the classification accuracy. In this paper the effect of the dimension reduction by the principal component analysis on the classification accuracy is experimentally studied.

1.4 Comparative study on statistical classification techniques

In order to evaluate the efficiency of the variable transformation and the principal component analysis, five classification techniques based on the Euclidean distance, Fisher's linear discrimination function, projection distance, modified projection distance[4] and the support vector machine (SVM) are employed in the classification test for the English text collection (the reuters-21578 [5][6]).

2 Procedure of Classification

The procedure of the automatic text classification consists of four general steps for feature vector generation, dimension reduction, learning and classification.

2.1 Feature vector generation

A feature vector for a text is composed of n feature elements each of which represents the frequency of a specific word in the text. At first a lexicon consisting of the all different words in a learning text set is generated. Then the feature vector for a text is composed of the frequencies of the lexicon words in the text. The dimensionality of the feature vector is equal to the lexicon size and is denoted by n. The normalization to the relative frequency is easily performed by (1), and the power transformation by (2).

2.2 Dimension reduction

At first the total covariance matrix of the learning sample is calculated to find the eigenvalues and eigenvectors. Each featurevector is transformed to the principal components in terms of the orthonormal transformation with the eigenvectors as the basis vectors. To reduce the dimensionality of the feature vector the principal components which correspond to the m largest eigenvalues are selected to compose the feature vector of dimensionality m ($< n$).

2.3 Learning

Parameters of each classification technique are determined in the training process using the learning sample. The Euclidean distance classifier employs the mean vector of each class. The linear discriminant function employs the weight vector determined by the mean vector of each class and the pooled within covariance matrix of entire classes. The projection distance (and the modified projection distance)employ the eigenvectors (and the eigenvalues) of the individual covariance matrix. As a support vector machine (SVM), C-support vector classification method (C-SVC) of linear type and of RBF type (with radial basis function) [7] were employed for the classification tests. We used the SVM library (LIBSVM Version 2.33) developed by Chang and Lin (2002) [8].

2.34 Classification

The feature vector of reduced dimensionality is classified to the class the distance (or the discriminant function) of which is minimized. Referring to the subject field manually given to each article in Reuters-21578, the classification rate R is calculated by

$$R = \frac{x}{(x+y)} \times 100 \tag{3}$$

where x is the number of articles correctly classified, and y is incorrectly classified respectively.

3 Classification Experiments

Classification experiments were performed to comparatively evaluate the feature extraction and classification techniques using the data collection Reuters-21578, which is composed of 21578 articles manually classified to 135 categories. In the experiments total of 750 articles, 150 articles/category randomly selected from five categories (acq, crude, earn, grain, trade), were used.

Table 1. Classification rate (%) at the optimal feature dimensionality

Classifier	Absolute frequency		Relative frequency	
	Without power transformation	With power transformation	Without power transformation	With power transformation
Euclidean distance	73.7	87.9	87.5	90.9
Linear discriminant function	90.5	94.9	93.3	95.3
projection distance	90.1	92.0	94.1	95.3
Modified projection distance	92.1	93.1	94.9	95.2
SVM-Linear	90.3	94.0	92.9	94.3
SVM- RBF	92.3	92.1	94.3	94.3

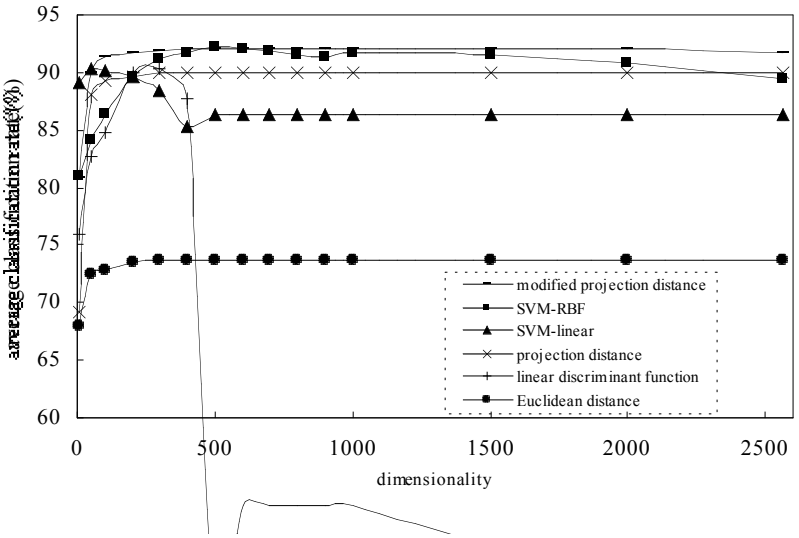

Fig. 1. Classification rate vs. dimensionality (absolute frequency)

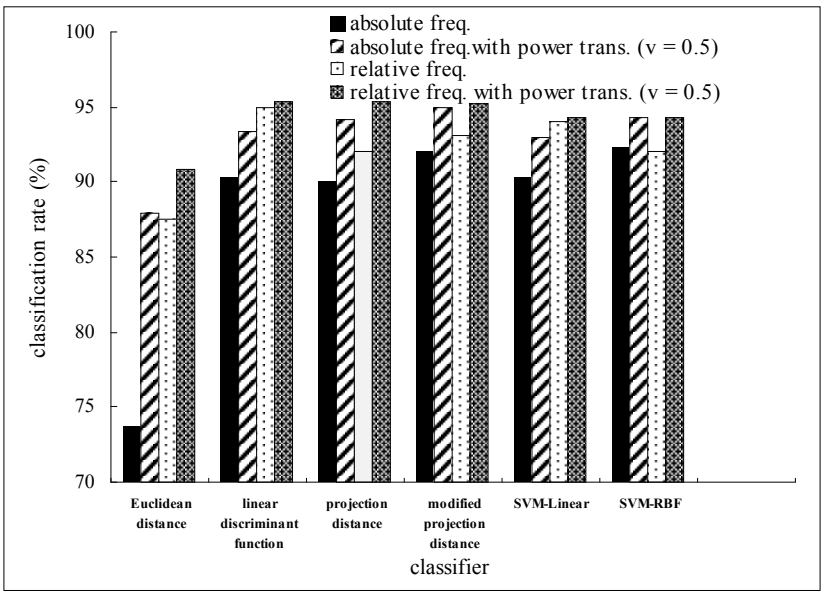

Fig. 2. Classification rate at the optimal feature dimension

Fig.1 shows the relationship between the average classification rate and the dimensionality of the feature vector composed of the absolute word frequencies. Table 1 and Fig.2 shows the classification rate at the optimal feature dimensionality.

The results are summarized as follows.

1. The classification rate was not sacrificed by the dimension reduction when the dimensionality was reduced to 15% (300-400 dim.) by the principal component analysis. Except for the linear discriminant function and the SVM-RBF the classification rate was not deteriorated significantly even when the dimensionality was further reduced to 5% (100 dim.).
2. The best classification rate was achieved by the linear discriminant function for small dimensionality (less than 50) and was achieved by the SVM-RBF for dimensionality from 450 to 600. The classification rate of the modified projection distance was totally the best for different dimensionality.
3. The classification rate was significantly improved by employing the relative frequency instead of the absolute frequency. The classification rate of the Euclidean distance classifier was most significantly improved from 73.7% to 87.5%.
4. The power transformation further improved the performance of each classification technique. When the power transformed relative frequency was employed the classification rate was over 94% for all classification techniques except for the Euclidean distance classifier.

4 Multi-classifier combination

Multi-classifiers were combined by majority vote to improve the classification accuracy. In this experiment the projection distance, the linear discriminant function and the SVM-linear were used to classify the feature vector, the component of which is the power transformed relative frequency. The final classification was performed by the majority vote of these three classifiers.

Table2 and Fig.3 show the result of classification test.

Table 2. Classification rate(%) of individual and combined classifiers

	Group1	Group2	Group3	average
linear discriminant function	93.6	96.4	96.0	95.3
projection distance	93.6	96.8	95.6	95.3
SVM-linear	92.4	96.4	95.2	94.6
Multi-classifier combination	93.2	97.6	96.4	95.7

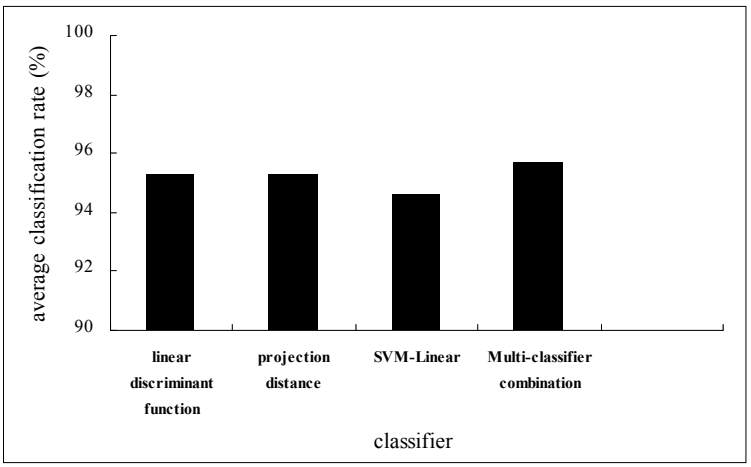

Fig. 3. Classification rate of individual and combined classifiers

Based on these results it is know that the classification rate was improved 0.4% using the multi-classifier combination.

5 Conclusions

This paper described a comparative study on techniques of feature transformation and classification to improve the accuracy of automatic text classification. The normalization to the relative word frequency, the principal component analysis (K-L transformation) and the power transformation were applied to the feature vectors, which were classified by the Euclidean distance, the linear discriminant function, the projection distance, the modified projection distance and the SVM.

The result of the experiments showed that

1. the principal component analysis drastically reduced the feature dimensionality without sacrificing the classification performance,
2. the normalization to the relative frequency followed by the power transformation improved the classifier performance significantly, and
3. considerably high classification rate for the transformed features was achieved by the linear discriminant function with less computational cost.
4. The classification rate can be improved using the multi-classifiers combination.

Intensive experimental evaluation employing more text samples of more categories is remaining as a future study. In order to simplify and clarify the performance evaluation, it was assumed that each article belonged to one category indicated by the first label in the subject list. The classification problem of multiply labeled articles is also remaining as a future study.

References

[1] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, 1-47, March 2002.

[2] Lam,W., Han,Y.: Automatic Textual Document Categorization Based on Generalized Instance Sets and a Metamodel. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.25, No.5, 628-633, May 2003.

[3] Fukunaga,K.: Introduction to Statistical Pattern Recognition, 76-77, Academic Press, Inc , 1990.

[4] Fukumoto,T., Wakabayashi,T. Kimura,F. and Miyake,Y.: Accuracy Improvement of Handwritten Character Recognition By GLVQ, Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition Proceedings(IWFHR VII), 271-280 September 2000.

[5] Sebastiani,R. Sperduti,A. and Valdambrini,N.: An Improved boosting algorithm and its application to text categorization, Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM 2000), 78-85 ,2000

[6] Yang,Y. and Liu.X. :A re-examination of text categorization methods, Proceedings of the Twenty-First International ACM SIGIR Conference on Research and Development in Information Retrieval, 42-49 ,1999

[7] Cortes,C. and Vapnik,V. : Support-vector network, Machine Learning 20, 273-297 ,1995.

[8] Chang,C.C. and Lin,C.J. : LIBSVM -- A Library for Support Vector Machines （Version 2.33）, http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html,.