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.

C.-H. Chi and K.-Y. Lam (Eds.): AWCC 2004, LNCS 3309, pp. 463–468, 2004.

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