Abstract

Text categorization (TC) is the task of automatically organizing a set of documents into a set of pre-defined categories. Over the last few years, increased attention has been paid to the use of documents in digital form and this makes text categorization becomes a challenging issue. The most significant problem of text categorization is its huge number of features. Most of these features are redundant, noisy and irrelevant that cause over fitting with most of the classifiers. Hence, feature extraction is an important step to improve the overall accuracy and the performance of the text classifiers. In this paper, we will provide an overview of using principle component analysis (PCA) as a feature extraction with various classifiers. It was observed that the performance rate of the classifiers after using PCA to reduce the dimension of data improved. Experiments are conducted on three UCI data sets, Classic03, CNAE-9 and DBWorld e-mails. We compare the classification performance results of using PCA with popular and well-known text classifiers. Results show that using PCA encouragingly enhances classification performance on most of the classifiers.
References


Index Terms

Computer Science Information Sciences

Keywords

Text Categorization, Dimension Reduction, Feature Extraction, Principle Component Analysis, Classifiers