ABSTRACT

Classification of text documents is a simple problem but very important issue because the benefits are quite large considering the number of documents every day is increasing. However, most existing document classification techniques require large amounts of labeled documents to perform training and testing phases. In classification of documents, on this final project used Principal Component Analysis algorithm combined with Support Vector Machines for supervised document. Principal Component Analysis is a technique that can be used to extract the structure of a high-dimensional data without losing any significant information to the whole data. Then it takes an algorithm that can generate prediction and accuracy of the document that is Support Vector Machines (SVM). SVM is a learning machine method that works on the principle of Structural Risk Minimization (SRM) in order to find the best hyperplane that separates the two classes in the input space. Hyperplane the best separator between the two classes can be found by measuring the margins of the hyperplane and look for the maximum point.

The results of system testing using data reduced by Principal Component Analysis (PCA) have slightly lower accuracy in certain dataset compared to without using PCA. The data that used is R8 data from Reuters-21578 Text Categorization of Collection Data Set. The best accuracy in this research was made by SVM method with average of 98.95%, while for SVM + PCA method which average accuracy 96.7866%.

Keywords :
Document Classification, Principal Component Analysis, Support Vector Machine