Abstract

Classification is the process of grouping text documents into different classes, in stages where each document is point to any particular class and require process to dig information from the document. Text preprocessing that will be done in this thesis include techniques of retrieving data such as word tokenization, feature selection, and term weighting to represent documents in the Vector Space Model. In feature selection, it will be calculate the weights of all keywords then terms that larger than the limit threshold will be taken. To calculate the weight of keywords, there are several methods that can be used such as Term Frequency (TF), TF * IDF (Inverse Document Frequency) and Information Gain (IG). Conducting the weighting terms, the concept of TF and TFIDF weighting calculation is used with the additional normalization techniques such as L1 and L2.

To improve performance and efficiency of the classification of text documents, especially in multi-class classification with a large number of categories can be done by using error-correcting output codes (ECOC). ECOC is multi-class technique that reduces multi-class problem to a set of binary classification and combining the results of binary classification to predict class/multiclass label.

In testing results show that, first, the incorporation method Support Vector Machine (SVM) with Error Correcting Output Coding (ECOC) can improve accuracy when compared with the method of Support Vector Machine (SVM) as the number of data used in 10%, 25%, 75% and 100% of the amount of data in each category. Second, the incorporation method Support Vector Machine (SVM) with Error Correcting Output Coding (ECOC) can improve accuracy when compared with the method of Support Vector Machine (SVM) at its threshold value greater than 0.03. Third, the merger method of Support Vector Machine (SVM) with Error Correcting Output Coding (ECOC) can improve accuracy when compared with the method of Support Vector Machine (SVM) when use Term weighting without normalization and fourth, fusion methods of Support Vector Machine (SVM) with Error Correcting Output Coding (ECOC) can improve accuracy when compared with the method of Support Vector Machine (SVM) when using SVM parameters with C value is large, kernel polynomial with small degree and RBF kernel with a large gamma value.

Keywords: Classification, Support Vector Machine, ECOC, *preprocessing*, accuracy