Abstract

A feature selection algorithm with redundancy reduction for text classification

Document classification involves the act of classifying documents according to their content to predefined categories. One of the main problems of document classification is the large dimensionality of the data. To overcome this problem, feature selection is required which reduces the number of selected features and thus improves the classification accuracy. In this paper, a new algorithm for multi-label document classification is presented. This algorithm focuses on the reduction of redundant features using the concept of minimal redundancy maximal relevance which is based on the mutual information measure. The features selected by the proposed algorithm are then input to one of two classifiers, the multinomial naive Bayes classifier and the linear kernel support vector machines. The experimental results on the Reuters dataset show that the proposed algorithm is superior to some recent algorithms presented in the literature in many respects like the F1-measure and the break-even point.