Abstract

Document categorization is one of problem in text mining. Classification technique is one of ways to categorize the document. Documents not only have high dimension of feature space, but also can have imbalance data characteristic. This imbalance will reduce the accuracy of data classification which is going to be built. One of solutions to increase the efficiency and the accuracy in classification document is by using feature selection technique.

This final project do the comparison analysis feature selection methods, such as Odds Ratio (OR), GSS Coefficient, Information Gain (IG), improved OR (iOR), and improved SIG (iSIG). These feature selection methods are implemented in filter feature selection, whereas for wrapper feature selection implement Odds Ratio (OR). Implementation use multinomial naive bayes classification technique. The method use naive bayes alghorithm which for calculate upon amount of words that appear in document. Beside using multinomial naive bayes, Implementation in filter feature selection also use the process of document classification which available in software Weka 3.5. By using the comparison analysis feature selection methods, it find what method that the most reliable to handle the imbalance data by testing the accuracy level data after being classified by test set. Data that is used comes from Reuters 21578 that imbalace characteristic.

Keywords : text mining, classification, imbalance, feature selection, multinomial naive bayes