Abstract

Text classification has become much more relevant with the increased volume of unstructured data from various sources. Several techniques have been developed for text classification. High dimensionality of feature space is one of the established problems in text classification. Feature selection is one of the techniques to reduce dimensionality. Feature selection helps in increasing classifier performance, reduce over filtering to speed up the classification model construction and testing and make models more interpretable. This paper presents an empirical study comparing performance of few feature selection techniques (Chi-squared, Information Gain, Mutual Information and Symmetrical Uncertainty) employed with different classifiers like naive bayes, SVM, decision tree and k-NN. Motivation of the paper is to present results of feature selection methods on various classifiers on text datasets. The study further allows comparing the relative performance of the classifiers and the methods.

References

38. 7 (2011): 8170-8177.
- DING, Xiaoming, and Yan TANG. "Improved Mutual Information Method For Text Feature Selection." 

**Index Terms**

Computer Science

Information Sciences

**Keywords**

Feature Selection  Filter Method  High Dimensionality  Text Classification  Text Categorization.