Examination of text categorization methods on open source classification toolkits

Hui Zhang
School of Library and Information Science

Abstract

This study compares two open source classifiers, Rainbow (McCallum, 1998) and Weka (Witten & Frank, 2005), by examining the performance of three text categorization methods: Naïve Bayes (NB), k-Nearest Neighbor (kNN), and the Support Vector Machine (SVM). Although there are extensive researches on evaluating and comparing the performance of different classifiers (Yang & Liu, 1999) and feature selection methods (Yang & Pedersen, 1997), studies of open source classifier evaluation are rarely published. Consequently, the researcher who wants to use an open source tool in his/her project has little information with which to make an appropriate choice of the classification software. This study focuses on the practical questions that are important to such a user group, which are listed below:

- What are the capabilities and limitations of open source classification toolkits?
- Do open source classification tools yield the same conclusions on feature selection methods and classifier performances as reported in the text categorization literature?
- Are there performance differences among open source toolkits?

Using accuracy, precision, recall and F1 measures with macro- and micro-averaging, this study will evaluate the classifier performances of Rainbow and Weka and compare them with the benchmark reported in Sebastiani' survey paper on text categorization (Sebastiani, 1999). In the context of text categorization, accuracy measures the classifier's ability to correctly identify true positive and true negative instances, and

F1 combines precision and recall to balance out the bias of single measure. In multi-class categorization, both macro- and micro-averaging methods are applied since micro-averaging "tends to be dominated by the classifier's performance on common categories", and the macro-averaging "tends to be more influenced by the performance on rare categories" (Yang & Liu, 1999).

Two tasks are assigned to each toolkit: the first task is to classify news articles of the Reuters21578 corpus with modified Apte split, and the second task is to identify spam emails in the TREC-2005 SPAM corpus. The first task investigates whether or not open source classifiers yield the same conclusions and performance level as reported in the literature, and the second task assesses the capacity and limitation of the toolkits by appliying TREC SPAM filtering requirement to classifiers. In addition to the two tasks described, comparison between two toolkits is conducted by text categorization methods.

Reference:

McCallum, A (1998). Rainbow. Retrieved July 26, 2005 from Carnegie Mellon University, School of Computer Science Web site: http://www-2.cs.cmu.edu/~mccallum/bow/rainbow/

Sebastiani, F., Machine learning in automated text categorization, Tech. Rep. IEI-B4-31-1999, Consiglio Nazionale delle Ricerche, Pisa, Italy, 1999.

Yang, Y., Pedersen, J.O., A Comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning ICML97, pp. 412---420, 1997.

Yang, Y., and Liu, X. A re-examination of text categorization methods. In SIGIR-99, 1999.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.