Improving Techniques for Naïve Bayes Text Classifiers

Han-joon Kim
University of Seoul, Korea

ABSTRACT

This chapter introduces two practical techniques for improving Naïve Bayes text classifiers that are widely used for text classification. The Naïve Bayes has been evaluated to be a practical text classification algorithm due to its simple classification model, reasonable classification accuracy, and easy update of classification model. Thus, many researchers have a strong incentive to improve the Naïve Bayes by combining it with other meta-learning approaches such as EM (Expectation Maximization) and Boosting. The EM approach is to combine the Naïve Bayes with the EM algorithm and the Boosting approach is to use the Naïve Bayes as a base classifier in the AdaBoost algorithm. For both approaches, a special uncertainty measure fit for Naïve Bayes learning is used. In the Naïve Bayes learning framework, these approaches are expected to be practical solutions to the problem of lack of training documents in text classification systems.

INTRODUCTION

The Naïve Bayes learning based classifier is simple yet surprisingly accurate, and thus has been used in many machine-learning related classification projects (A-Engelson & Dagan, 1999; Katakis, Tsoumakas & Vlahavas, 2006; Nigam, McCallum, Thrun & Mitchell, 1998). In particular, it is remarkably successful for text classification problems, despite the fact that text data generally has a large feature space (Katakis, Tsoumakas & Vlahavas, 2006; Nigam, McCallum, Thrun & Mitchell, 1998). Compared to other learning methods, the Naïve Bayes has a number of good features that serve text classification systems.
• The learning of the Naïve Bayes classifier does not demand any other statistics than feature statistics, and no further complex generalization processes are required unlike the other machine learning methods such as support vector machine.

• It is very easy to incrementally update the classification model of a given categories due to its simplicity. When a new training documents arrive, the feature statistics are updated and feature evaluation can be immediately calculated without the need of re-processing past data. This characteristic is essential in the case where the document collection is highly evolutionary.

• Since a classification model can be developed with a single pass over the documents, a Naïve Bayes learning process is faster than that of other methods.

• The Naïve Bayes learning can easily accommodate the degree of importance of features occurring in documents; for example, for learning, one may double or triple the frequency of terms occurring in titles in news articles and title tags in HTML documents.

Because of the above advantages, there has been a strong incentive to improve the Naïve Bayes text classifier by combining it with other learning techniques (Kim & Kim, 2004; Kim, Rim, Yook & Lim, 2002; Nigam, McCallum, Thrun & Mitchell, 1998). To improve the classification performance, we first should resolve the problem of lack of training examples in the real-world environment. In general, most machine learning methods including Naïve Bayes assumes the existence of good quality documents for training. However, this assumption is not effective in real world operational environments. Thus we must assume that rather than trying to prepare complete training examples at one time, new training documents are continuously provided for learning as a data stream. Thus in such an environment, continuous update of the current classification model is required whenever a new set of training examples are prepared. Thus, how to obtain training examples has become an important issue in practically developing a text classifier. To resolve such a problem, active learning approach was proposed in which the learner can actively choose the training documents from a pool of unlabeled documents (A-Engelson & Dagan, 1999). However, active learning approach needs additional cost for labeling unknown documents by human experts. In this chapter, I will introduce two approaches to improving Naïve Bayes with incomplete training set: one is the EM (Expectation Maximization) approach and the other is the Boosting one.

First, EM approach is a very successful solution to the problem of difficulty of obtaining training documents. EM algorithm is an iterative method for finding maximum likelihood in problems with incomplete (or unlabeled) data. It allows learning more accurate text classifiers from only a few labeled examples by augmenting the labeled documents with a large pool of unlabeled documents. Currently, EM approach has been evaluated to be a practical and excellent solution to the problem of lack of training examples in text classification systems (Nigam, McCallum, Thrun & Mitchell, 1998) Naïve Bayes in combination with EM algorithm dramatically improved overall accuracy of text classification with only a few labeled training examples (Nigam, McCallum, Thrun & Mitchell, 1998). Second, Boosting is a meta-learning technique for improving the accuracy of machine-learning based classifiers, which combines a series of base (or weak) classifiers to produce a single powerful classifier. Since the Boosting technique was developed by Freund and Schapire (1996), it has been considered to be one of the best approaches to improving classifiers in many previous studies (Schapire & Singer, 2000). In particular, Boosting contributes to significantly improve the decision tree learning algorithm (Freund & Schapire, 1996; Friedman, Hastie & Tibshirani, 2000; Quinlan, 1996).