Techniques for Improving the Performance of Naive Bayes for Text Classification

Karl-Michael Schneider

University of Passau, Department of General Linguistics
Innstr. 40, 94032 Passau, Germany
schneide@phil.uni-passau.de
WWW home page: http://www.phil.uni-passau.de/linguistik/schneider/

Abstract. Naive Bayes is often used in text classification applications and experiments because of its simplicity and effectiveness. However, its performance is often degraded because it does not model text well, and by inappropriate feature selection and the lack of reliable confidence scores. We address these problems and show that they can be solved by some simple corrections. We demonstrate that our simple modifications are able to improve the performance of Naive Bayes for text classification significantly.

1 Introduction

Text classification is the assignment of predefined categories to text documents. Text classification has many applications in natural language processing tasks such as E-mail filtering [1, 2], news filtering [3], prediction of user preferences [4] and organization of documents [5]. Because of the variety of languages, applications and domains, machine learning techniques are commonly applied to infer a classification model from example documents with known class labels. The inferred model can then be used to classify new documents. A variety of machine learning paradigms have been applied to text classification, including rule induction [6], Naive Bayes [7], memory based learning [8], decision tree induction [9] and support vector machines [10].

This paper is concerned with the Naive Bayes classifier. Naive Bayes uses a simple probabilistic model that allows to infer the most likely class of an unknown document using Bayes' rule. Because of its simplicity, Naive Bayes is widely used for text classification [4, 5, 1, 2, 11].

The Naive Bayes model makes strong assumptions about the data: it assumes that words in a document are independent. This assumption is clearly violated in natural language text: there are various types of dependences between words induced by the syntactic, semantic, pragmatic and conversational structure of a text. Also, the particular form of the probabilistic model makes assumptions about the distribution of words in documents that are violated in practice [12]. Nonetheless, Naive Bayes performs quite well in practice, often comparable to more sophisticated learning methods [13, 14].

One could suspect that the performance of Naive Bayes can be further improved if the data and the classifier better fit together. There are two possible approaches: (i) modify the data, (ii) modify the classifier (or the probabilistic model).