UNIVERSITY OF TEXAS AT DALLAS
DEPARTMENT OF COMPUTER SCIENCE

Exploring the Techniques to Improve the Performance of Naïve Bayes for Spam Email Filtering and Email Categorization

Chao Wang and Chao Li

November 9, 2013

1 PROBLEM DEFINITION

In the domain of personal email messages, text categorization methods have been widely applied to the problem of spam filtering and email categorizing into folders. The spam filtering task is a two-value classification: spam and non-spam, while the email categorizing is considered as a multi-class categorization problem like what Gmail does.

Naïve Bayes classifier is a simple generative classification algorithm that has reasonably good performance. But it makes strong assumption that all words in a document are independent. This assumption is clearly lack of context information which violates natural language rules [5]. Therefore, Naïve Bayes classifier needs some modification to improve its performance.

In our final project, we seek to implement and compare some existing approaches which can improve the performance of Naïve Bayes in spam email filtering and email categorization. Through such implementation and comparison, we hope we can find our own ways to improve the classification accuracy by doing some modification to Naïve Bayes.

2 PLAN AND METHODS

We have checked many papers about improving the Naïve Bayes method in text classification, and choose four methods as our mainly referred methods: Locally weighted naive bayes [4],

Adjusted probability naive Bayesian [6], Learning weighted naive Bayes [7] and some Naïve Bayes improvement method without specific name [5]. These methods are all widely applied in many natural language processing fields. We plan to implement three to four methods among them, and apply these methods in spam email filtering and email categorization. Meanwhile, we will also try to explore a new method to improve the traditional Naïve Bayes classifier in our experiment. As to the results, we seek to obtain the accuracy result and precision-recall curve of these methods on some common widely used database.

3 Resources

We plan to use two database in our project: Apache Spam email data [3] by Apache Spam assassin project, and Enron Email Dataset [2] by CMU. The first one is for spam filtering and the latter one is for email categorizing. In particular, Ref. [1] utilizes many categorization methods including Naïve Bayes on the Enron Email Dataset. Therefore, it is straightforward for us to compare the results on the Enron Dataset utilizing results of Ref. [1].

References

[1] Ron Bekkerman. Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. 2004.

[2] Enron Email Dataset. `http://www.cs.cmu.edu/~enron/`.

[3] Apache Spam email dataset. `http://spamassassin.apache.org/publiccorpus/`.

[4] Eibe Frank, Mark Hall, and Bernhard Pfahringer. Locally weighted naive bayes. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 249–256. Morgan Kaufmann Publishers Inc., 2002.

[5] Karl-Michael Schneider. Techniques for improving the performance of naive bayes for text classification. In *Computational Linguistics and Intelligent Text Processing*, pages 682–693. Springer, 2005.

[6] Geoffrey I Webb and Michael J Pazzani. Adjusted probability naive bayesian induction. In *Advanced Topics in Artificial Intelligence*, pages 285–295. Springer, 1998.

[7] Harry Zhang and Shengli Sheng. Learning weighted naive bayes with accurate ranking. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pages 567–570. IEEE, 2004.