INTEGRATING NAÏVE BAYESIAN CLASSIFIER WITH SUPPORT VECTOR MACHINE FOR SPAM CLASSIFICATION

Chui-Yu Chiu, Chienwen Wu, Yuan-Ting Huang,
Department of Industrial Engineering and Management, National Taipei University of Technology
1,Sec.3,Chung-Hsiao E. Rd.,Taipei, Taiwan, R.O.C., 886-2-27712171 Ext.2365
Email:cychiu@ntut.edu.tw

ABSTRACT

In this research we propose a spam classification method which integrates the Naïve Bayesian Classifier (NBC) and Support Vector Machine (SVM). NBC adopts the concept of Bayesian theory for classification, and combines the conditional probability with feature count as input data for SVM. The classification features generated from spam data set are used to train and test the proposed method.

INTRODUCTION

To deal with the problems of spam, many approaches such as decision trees, k-nearest neighbor, back-propagation neural network, Bayesian methods and support vector machines have been proposed. Naïve Bayesian Classifier (NBC) is a simplified form of Bayes' rule that assumes independence of the observations. Researches [3][9] demonstrated that NBC has competitive performance in comparison with other learning algorithms. On the other hand, Support Vector Machine (SVM) is one of the most widely used machine learning techniques for classification and regression developed by V. Vapnik, and has better performance in text classification, pattern segment and spam classification, etc. [6][11][12][15].

For text classification, features which consist of words or symbols are selected to represent the entire text or mail. In the first stage of this research, NBC is proposed to classify data set and calculate the conditional probability of features generated from the procedures of feature selection. In the second stage of this research, SVM is applied to improve the classification performance. Eventually, the proposed two-stage classifier, NBC+SVM, is compared with other known classification methods using the precision and recall rate.

THE PROPOSED CLASSIFIER NBC+SVM

Feature Selection

In English texts or mail, words are explicitly separated by white spaces, and extraction of words is straightforward. Furthermore, feature selection is performed to choose representative terms for each class such that these terms can distinguish one class from the others. Some feature selection methods have better performance on text classification problems like Term Frequency- Inverse Document Frequency (TF-IDF), Information Gain, X^2 statistic, Mutual Information, etc,. As revealed in previous research [14], TF-IDF is a well-known methods with better performance, thus we apply the TF-IDF to feature selection.

Integration of NBC and SVM

The feature counts are usually treated as the input data for SVM to train the classifier. However, the reason for using the feature count as input data to SVM may be insufficient in spam classification. Therefore, in our proposed two-stage classifier, we apply not only the concept of feature count but also the conditional probability of features from the first-stage classifier NBC[1][2][7]. In other words, the values of feature counts are multiplied by the conditional probability of NBC to represent the input data of SVM. The input data of SVM is composed of the conditional probability and feature count.
We describe the proposed two-stage classification method as following steps.

1 Prepare the spam data set (Ling-Spam) for the proposed classifier.
2 Select features from the prepared spam data set by TF-IDF for feature selection.
3 Calculate the prior and conditional probability (representative probability of features) of NBC as the first stage of proposed classifier.
4 Calculate the posteriori probability of NBC and classify the spam data set (Ling-Spam).
5 Combine the conditional probability of NBC and feature count as input data to train SVM classifier.
6 Combine the conditional probability of NBC and feature count as input data to test SVM classifier.
7 Calculate the precision and recall rate from the evaluated results and compare with other methods.

EXPERIMENT PROCESS AND RESULTS

In first-stage of this research, we use Visual Basic 6.0 to develop the program of NBC. Furthermore, the probability values of features are inputted to the Access data file format and connected to the primary program. When finishing the NBC in the first-stage classification, SVM is proposed to the second-stage classification. In this stage, we use C.J. Lin's LIBSVM [5] for our experiment. Table 1 is the results of precision and recall rate of NBC+SVM classifier. According to the result, NBC+SVM classifier has best performance compared with other classification methods.

Table 1. The precision and recall rate of NBC+SVM and other methods

Method Criterion	**NBC+SVM**	**Memory Based k-NN [10]**	**Outlook Pattern**	**Boosting Tree[4]**
Precision Rate (%)	96.18	95.62	87.93	91.27
Recall Rate (%)	93.39	85.27	53.01	84.80

CONCLUSIONS

In this research, we propose a NBC+SVM classifier for the spam classification problem. Results show that NBC+SVM classifier has better precision and recall rate than other classification methods.

ACKNOWLEDGEMENTS

This work has been supported by the National Science Council of Taiwan, under project NSC 94-2213-E-027-008

REFERENCES

[1] Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Paliouras, G., and Spyropoulos., "An evaluation of Naïve Bayesian anti-spam filtering," In *Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)*, Barcelona, Spain, pp. 9–17, 2000.

[2] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C.D., and Stamatopoulos, P., "Learning to filter spam e-mail: a comparison of a naïve Bayesian and a memory-based approach". In *Proceedings of the Workshop on Machine Learning and Textual Information Access, PKDD 2000*, Lyon, France, pp. 1– 3.

[3] Buntine, W., "Learning classification rules using Bayes," *Proc. 6th Int. Workshop Machine Learning*, pp. 94–96, 1989.

[4] Carreras, X. and Marquez, L., "Boosting trees for anti-spam e-mail filtering," In *Proceedings of the 3rd Conference on Recent Advances on NLP*, RANLP '01, 2001.

[5] Chang, C. C. and C. J. Lin., "LIBSVM: a library for support vector machines," *Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm*, 2001.

[6] Drucker, H., Wu, D. and Vapnik, VN., "Support. Vector Machines for Spam Categorization," *IEEE. Transactions on Neural Networks*, Vol. 20, No. 5, Sep. 1999.

[7] Ganapathiraju, A., Hamaker, J.E. and Picone, J., "Applications of support vector machines to speech recognition Signal processing," *IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, Volume 52, Issue 8, Page(s):2348 – 2355, 2004.

[8] Guthrie, L. and Walker, E., "Some comments on document classification by machine," Unpublished, 2000.

[9] Langley, P. W., Iba, and Thompson, K., "An analysis of Bayesian classifiers," *Proc. 11th National Conference of Artificial Intelligent*, pp. 223–228, 1992. [10]Lippmann, Richard P., "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, 4-22, April 4, 1987.

[11] Li, K. L., Li, Kai., Huang, H. K. and Tian, S. F., "Active Learning with Simplified SVMs for Spam Categorization," *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference* 3, Nov, 4-5, 2002.

[12] Mao, K.Z., "Feature subset selection for support vector machines through discriminative function pruning analysis," *Systems, Man and Cybernetics, Part B, IEEE Transactions* 34 (1), Feb, 60 – 67, 2004.

[13] Oda, T. and White, T., "Increasing the accuracy of a spam-detecting artificial immune system," *Evolutionary Computation, CEC '03, The 2003 Congress on*, 1 (11), Dec, 390 – 396, 2003.

[14] Salton, G. and Buckley, C., "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, Vol. 24, No. 5, pp. 513-523, 1988. [15]Vapnik, V. N., "Statistical Learning Theory," New York: Wiley, 1998.