Journal of Computational Information Systems 8: 3 (2012) 949–956
Available at http://www.Jofcis.com

Spam Email Classification Using Decision Tree Ensemble

Lei SHI*, Qiang WANG, Xinming MA, Mei WENG, Hongbo QIAO

College of Information and Management Science, HeNan Agricultural University, Zhengzhou 450002, China

Abstract

Spam email has already caused many problems such as taking recipient time and wasting network bandwidth. It is time consuming and laborious to remove spam email by hand if there are too many spam email in mailbox. Thus, automatic classification of spam email from legitimate email has become very important. Decision tree and ensemble learning are two popular and powerful techniques in machine learning community. In this study, a novel classification method based decision tree and ensemble learning is introduced to classify the spam email effectively. An experimental evaluation of different methods is carried out on a public spam email dataset. The experimental results suggest that the proposed method generally outperforms benchmark techniques.

Keywords: Spam Email; Decision Tree; Ensemble Learning

1 Introduction

Spam email is an email sent to somebody without consent and its content can cause unease and distress [1]. Usually, the spam email is unsolicited and sent in bulk and the sender of spam email doesn't target recipients personally. Thus, the addresses of recipients often are guessed and the same spam email is sent to numerous people at the same time [1]. These spam emails have already caused many problems such as consuming network bandwidth, wasting recipient time and so on [2, 3]. To resolve these problems, classification of spam email from legitimate email has become very important. Recently, many machine learning and data mining techniques have been applied in spam email classification, such as Naive Bayes [4-6], Support Vector Machines (SVM) [7-9] and rule learning [10].

Decision tree is very popular and powerful tool in data mining community and the rules generated by decision tree are simple and accurate for most problems [11]. Ensemble learning is a novel technique where a set of individual classifiers are trained and jointly used to solve a problem [12]. The basic principle of ensemble learning is that no single classifier can claim to be uniformly superior to any other classifier. Because the combination of several component classifiers will enhance the accuracy and reliability of the final classifier, an ensemble classifier can have overall better performance than the individual component classifiers. Ensemble learning is attracting

[*] Corresponding author.
Email address: `sleicn@126.com` (Lei SHI).

February 2012

more and more attention from data mining and machine learning domains because of its good generalization. Many researches have shown that ensemble of classifiers is very effective for classification tasks and can be applied successfully in many different areas such as information retrieval and text classification [13]. In this study, a novel classification method based decision tree and ensemble learning is introduced to classify the spam email effectively.

This paper is organized as follows. Section 2 first introduces decision tree and ensemble learning briefly. Then, the proposed approach for classifying the spam email is introduces in detail. Section 3 presents and discusses the experimental results compared with existing approaches. Section 4 gives the conclusions and discusses the prospects for future work.

2 The Proposed Approach

2.1 Decision tree

The decision tree is one of the most famous tools of decision-making theory. Decision tree is a classifier in the form of a tree structure to show the reasoning process. Each node in decision tree structures is either a leaf node or a decision node. The leaf node indicates the value of the target attribute of instances. The decision node indicates two or more branches and each branch represents values of the attribute tested. When classifying an unknown instance, the unknown instance is routed down the tree according to the values of the attributes in the successive nodes. C4.5 is one of the most popular decision trees algorithms [11, 14]. According to the splitting node strategy, C4.5 builds decision trees from a set of training data. At each node of the tree, C4.5 chooses one attribute that most effectively splits its set of instances into subsets. The C4.5 algorithm recursively visits each decision node and selects the optimal split until no further splits are possible. Fig. 1 is an illustration of the structure of decision tree built by the C4.5.

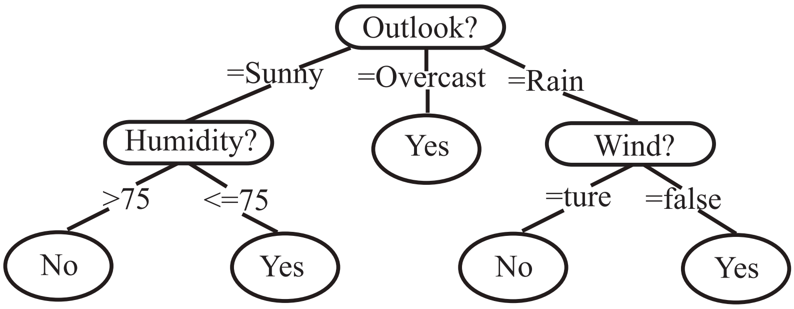

Fig. 1: An example of decision tree

In Fig.1, Outlook, Humidity and Wind in inner nodes of the tree are condition attributes and Yes and No are the values of decision attribute in the dataset.

2.2 Ensemble learning

A supervised machine learning algorithm is to construct a mapping from input data to the appropriate outputs. In recent years, ensemble learning has become one of the most important fields in the machine learning community. Compared with traditional machine learning techniques which generate a single classifier, ensemble learning methods instead generate multiple classifiers. Ensemble methods commonly comprise two main phases. The first phase of an ensemble method is to product different classifiers. The second phase of an ensemble method is to combine the classifiers by using some strategies such as weighted voting and stacking approaches.

2.3 Proposed approach

To improve the classification performance of spam email, the proposed approach uses C4.5 as base classifiers and applies ensemble learning technique to generate the final output by combining the predictions of them. The basic framework of the proposed decision tree ensemble based classification algorithm of spam email is shown in Fig. 2.

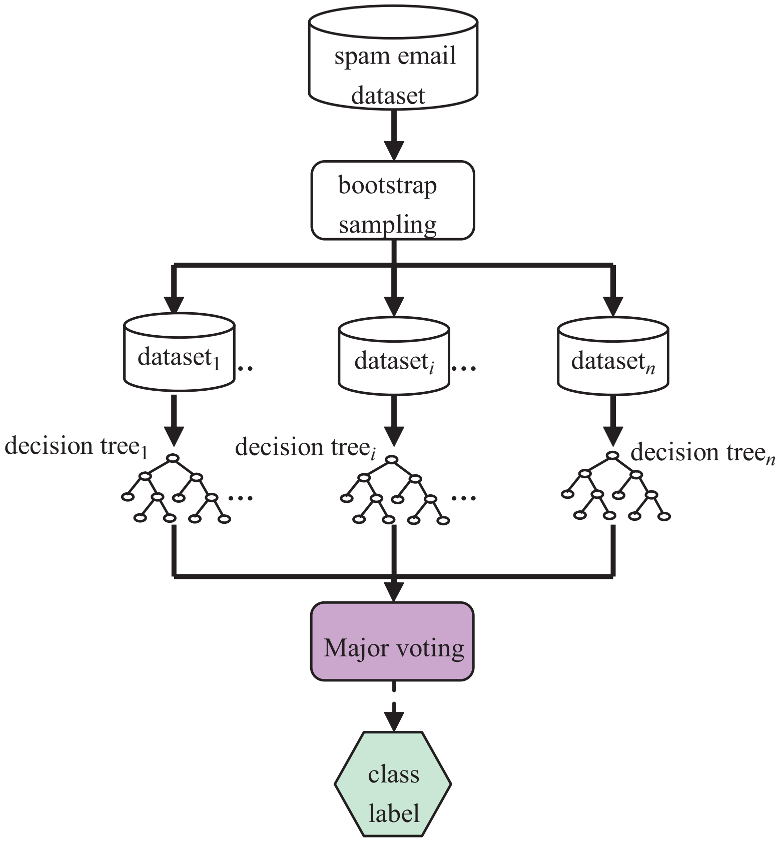

Fig. 2: The schematic view of the proposed algorithm

In the proposed method, each classifier's training set is generated by selecting instances at random with replacement from the original spam email training dataset and the the number of selected instances is the size of the original spam email training dataset. Thus, many of the original instances may be repeated in the resulting training set while others may be left out. Then, the component decision tree C4.5 classifier is training from classifier's training set and some C4.5 classifiers will be obtained. Prediction of a test instance by the proposed method is given by the uniform majority voting of component classifiers. The proposed method is described as Algorithm 1.

Algorithm 1. Decision tree ensemble based classification algorithm of spam email

Input: spam email training set L; size of ensemble T; decision tree algorithm C4.5.

```
Output: $C^*$
(1) For $i = 1$ to $T$
(2) Create a training set $L_i$ from the original spam email training set $L$. The size of $L_i$ is the same with the $L$ where some email instances may not appear in it while others appear more than once
(3) Build a decision tree classifier $C_i$ on the $L_i$
(4) End For
(5) The final classifier $C^*$ is formed by aggregating the $T$ decision tree classifiers
(6) Classify a new email $x$, a vote for class $y$ is recorded by every decision tree classifier $C_i(x) = y$
(7) $C^*(x)$ is the class with the most votes, i.e., $C^*(x) = argmax_y(\sum_{i=1}^{T} \psi(C_i(x) = y))$ ($\psi$ is an index function such that $\psi(true) = 1$, $\psi(false) = 0$.)
```

3 Experiments

In this section, experimental evaluation of the proposed method is reported. Firstly, spam email dataset used for evaluation purposes is described. Then, evaluation procedure and criteria is introduced. The experimental results and related discussion are presented finally.

3.1 Experimental datasets

The empirical evaluation employs a publicly available spam email dataset, namely, SPAM E-mail Database [15]. The SPAM E-mail Database contains 58 attributes and a total of 4601 emails, of which 1813 emails are spam emails and 2788 emails are normal emails.

3.2 Performance measures

In this work, we use the Accuracy, F_1 and AUC as the main methods for assessing our experiments. Four cases are considered as the result of classifier to the instance in Table 1 [16].

Table 1: Cases of the classification for one class

Class C		Result of classifier	
		belong	Not belong
Real classification	Belong	TP	FN
	Not belong	FP	TN

TP (True Positive): the number of instances correctly classified to that class.

TN (True Negative): the number of instances correctly rejected from that class.

FP (False Positive): the number of instances incorrectly rejected from that class.

FN (False Negative): the number of instances incorrectly classified to that class.

Accordingly, Accuracy, precision, recall and F_1 measure can be defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

$$precision = \frac{TP}{TP + FP}. \tag{2}$$

$$recall = \frac{TP}{TP + FN}. \tag{3}$$

$$F_1\ measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

The Accuracy of the classifier is the proportion of instances which are correctly classified. F_1 is the harmonic mean of precision and recall, where precision is the rate of instances classified correctly among the result of classifier and recall is the rate of correct classified instances among them to be classified correctly [16].

In the machine learning literature, the Area Under the receiver operating characteristic Curve (AUC) is a famous method for comparing classifiers, and it represents the expected performance as a single scalar [18]. The AUC has a known statistical meaning, i.e., it is equivalent to the Wilconxon test of ranks, and is equivalent to several other statistical measures for evaluating classification and ranking models [19].

3.3 Results and discussion

In the experiments, four popular machine learning algorithm C4.5, Naive Bayes, SVM and kNN are implemented for comparison. For the SVM, the LIBSVM [20] is used for implementation and radial basis function is set as default kernel function of SVM. For kNN, we set k=1. For the proposed algorithm, the number of bootstrap samples parameter T is set to 10. The statistics of classification performance is evaluated by 10-fold cross-validation approach to reduce the bias and variance of classification results. Each dataset is divided into 10 partitions and the algorithm is run once for each partition. Each time, nine partitions are grouped together for training and the remaining tenth is used for test. The training-test procedure is conducted ten times and the average of the ten performances is used as final result.

The Accuracy values of all techniques on the dataset are shown as Fig. 3.

According to the experimental results shown in Fig. 3, the proposed algorithm yields top-notch performance among these techniques. The Accuracy of proposed method is 94.6%, which is approximately 1.9% higher than that of C4.5, 15.4% higher than that of Naive Bayes, 11.1% higher than that of SVM, and 4.0% higher than that of kNN.

The F_1 values of all techniques on the dataset are shown as Fig. 4.

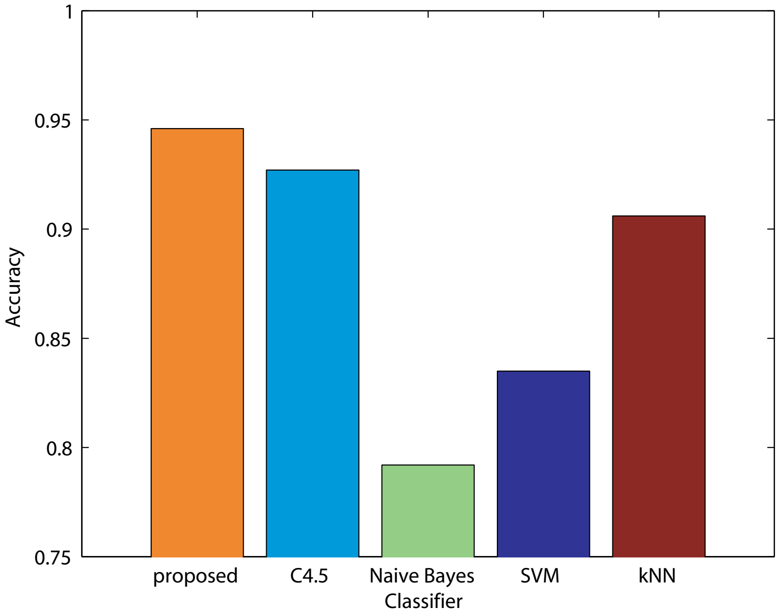

Fig. 3: Comparison of the accuracy of all techniques on the dataset

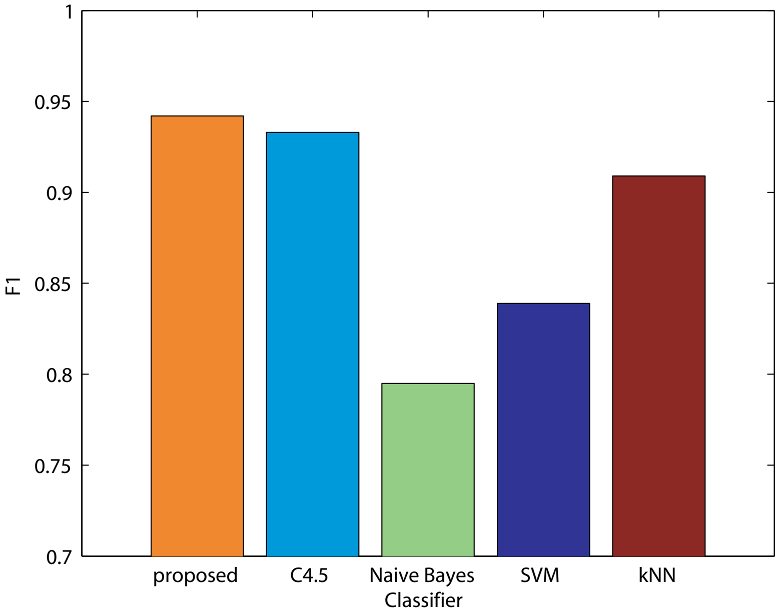

Fig. 4: Comparison of the F_1 of all techniques on the dataset

Fig. 4 indicates the proposed algorithm also yields a higher performance compared to other techniques. The F_1 of proposed method is 94.2%, which is approximately 0.9% higher than that of C4.5, 14.7% higher than that of Naive Bayes, 10.3% higher than that of SVM, and 3.3% higher than that of kNN.

The AUC values of all techniques on the dataset are shown as Fig. 5.

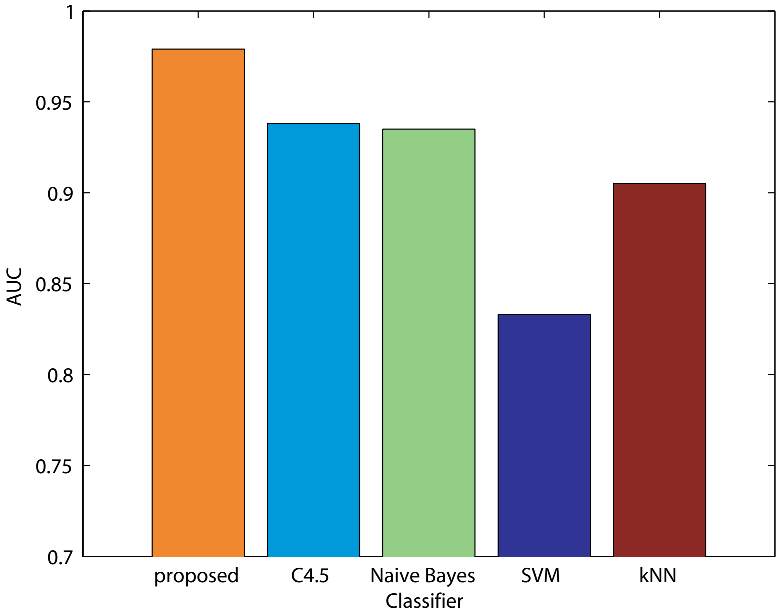

Fig. 5: Comparison of the AUC of all techniques on the dataset

According to the experimental results shown in Fig.5, the proposed method beats other algorithms by a wide margin. The AUC of proposed method is 97.9%, which is approximately 4.1% higher than that of C4.5, 4.4% higher than that of Naive Bayes, 14.6% higher than that of SVM, and 7.4% higher than that of kNN. Thus, experimental comparison clearly shows performance improvement of the proposed algorithm is significant.

4 Conclusion

In this paper, an ensemble learning and decision tree based approach is proposed to classify spam emails. Extensive experiments conducted on a public spam email dataset indicate that the proposed algorithm outperforms the popular classification techniques including of C4.5, Naive Bayes, SVM and kNN. In this paper, a monolingual spam email classification is researched. Our future work is to investigate the classification technique of multilingual spam email.

References

[1] http://en.wikipedia.org/wiki/E-mail_spam, 2011.

[2] D. Cook, Catching Spam before it arrives: Domain Specific Dynamic Blacklists, Australian Computer Society, ACM, 2006.

[3] R. Deepa Lakshmi et al, Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools, International Journal on Computer Science and Engineering 02(2010), 2760-2766.

[4] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos et al, An Evaluation of Naive 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, pp. 9-17.

[5] J. Hovold, Naive Bayes spam filtering using word-position-based attributes, in: Proceedings of the 2nd Conference on Email and Anti-Spam, 2005.

[6] V. Metsis, I. Androutsopoulos, G. Paliouras. Spam filtering with Naive Bayes-which Naive Bayes?, in: Proceedings of the 3rd Conference on Email and Anti-Spam, Mountain View, CA, 2006.

[7] H. Drucker, D.Wu, V. Vapnik, Support vector machines for spam categorization, IEEE Transactions on Neural Networks, 10 (5) (1999), 1048-1054.

[8] A. Kolcz, J. Alspector, SVM-based filtering of email spam with content-specific misclassification costs, in: Proceedings of Workshop on Text Mining, Chicago, IL, 2001.

[9] Q. Wang, Y. Guan, X.L. Wang, SVM-Based Spam Filter with Active and Online Learning, in: Proceeding of Text REtrieval Conference on Spam Filtering Task, 2006.

[10] W.W. Cohen, Learning Rules that Classify E-mail, in: Proceedings of AAAI Spring Symposium on Machine Learning in Information Access, 1996, pp. 18-25.

[11] J.R. Quinlan, C4.5:Programs for Machine Learning, Morgan Kaufmann, San Matteo, CA,1993.

[12] T. Dietterich, Ensemble methods in machine learning, in: Proceedings of First international workshop on multiple classifier systems, pp. 1-15.

[13] L. Shi, X.M. Ma, L. Xi, et al, Rough set and ensemble learning based semi-supervised algorithm for text classification, Expert Systems with Applications 38(5) (2011), 6300-6306.

[14] J.R. Quinlan, Bagging, boosting, and C4.5, in: Proc. Thirteenth National Conf. on Artificial Intelligence, 1996, pp. 725-730.

[15] w3.umh.ac.be/pub/ftp_ssi/teaching/dwdm/spambase.arff, 2011.

[16] Y. Yang, An evaluation of statistical approaches to text categorization, Information Retrieval (1999), 69-90.

[17] J. A. Hanley, Receiver operating characteristic (ROC) methodology: the state of the art, Critical Reviews in Diagnostic Imaging (1989), 307-35.

[18] S. Rosset, Model selection via the AUC, in: Proceedings of the 21th International Conference on Machine Learning, Banff, Alberta, Canada, 1989, pp. 89-97.

[19] D.J. Hand, Construction and assessment of classification rules, Wiley, New York, 1997.

[20] http://www.csie.ntu.edu.tw/c̃jlin/libsvm/, 2011.