DATA MINING METHODS IN DETECTION OF SPAM

Dong-Her Shih, Hsiu-Sen Chiang, Chia-Shyang Lin, Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Douliu, Yunlin, Taiwan, R.O.C, {shihdh,g9023728,g9223728}@yuntech.edu.tw

ABSTRACT

We compare three data mining methods for spam filtering - Naïve Bayes, Fisher's method of probability combination, and ID3 decision tree. The experimental result shows that although there is no dominant algorithm to the spam problem, but generally, the decision tree has the best performance. And Fisher's method has better performance than Naïve Bayes in general.

INTRODUCTION

Spam problem has brought enormous cost for enterprises and users that use Internet. The problem is also getting worse and worse. The main reason of the spam problem is that the additional cost is so cheap for the spammers to send another hundred or even thousand of recipients; thus, they will target as many email addresses as they possibly can get. Spammers also carry out dictionary attacks to mail service providers. Many studies have been made to deal with this problem. Filtering is currently the most widely in use. We compared three data mining method that can be used in spam filtering: ID3 Decision Tree, Naive Bayes Filter, and Fisher's Probability Combination Method. The performance measuring result shows that ID3 decision tree has better performance in general. Next, we will provide a brief describe for these methods and the experiment result.

DATA MINING METHOD

We employed three different data mining methods to generate filter for spam detection. Each method has different features.

Naïve Bayes

We assume that there are similar contents in spam mail and could be differentiated from legitimate. Each mail is represented by a vector $\vec{x} =< x_1, x_2, ..., x_n >$, where $x_1, x_2, ..., x_n$ are the values of attributes $X_1, X_2, ..., X_n$. Following [3], binary attributes are employed, that is, when the mail has the specific word represented by X_i, $X_i = 1$, otherwise $X_i = 0$. Given the vector $\vec{x} =< x_1, x_2, x_3, ..., x_n >$ and class $k \in \{spam, legitismate\}$, we can calculate the probability that an email belongs to class *c* with Eq.

(1).

$$P(C=c \mid \overrightarrow{X}=\vec{x})=\frac{P(C=c)\cdot P(\overrightarrow{X}=\vec{x} \mid C=c)}{\sum_{k\in\{spam,legitimate\}} P(C=k)\cdot P(\overrightarrow{X}=\vec{x} \mid C=k)} \tag{1}$$

Due to the combinations of $\overrightarrow{X}$ are too many and there are also data sparseness problem, the probabilities $P(\vec{X} \mid C)$ are almost impossible to calculate. The Naïve Bayes filter use Eq. (2) to calculate $P(C=c \mid \overrightarrow{X}=\vec{x})$ under the assumption of $X_1, X_2, X_3, ..., X_n$ are conditionally independent:

$$P(C=c \mid \vec{X}=\vec{x})=\frac{P(C=c)\cdot \prod_{i=1}^{n} P(X_i=x_i \mid C=c)}{\sum_{k\in\{spam,legitimate\}} P(C=k) \cdot \prod_{i=1}^{n} P(X_i=x_i \mid C=k)} \tag{2}$$

Fisher's Probability Combination Method

In [1] proposed a Bayes-like method that can release the independent assumption through R. A. Fisher's probability combination method. For each word that appears in the training data, we calculate:

$$p'(w)=\frac{p(w \mid spam)}{p(w \mid spam)+p(w \mid legitimate)} \tag{3}$$

$$f(w)=\frac{(s\times x)+(n\times p'(w))}{s+n} \tag{4}$$

$p'(w)$ can be interpreted as the probability that a randomly choose email that contains w word will be spam. Due to the rare data problem in training set, we combine the value with weights through Eq (4) in which s is the strength we want to give from our background information, x is the probability we assume, and n is the the number of email that contains w. Reasonable starting points of x and w_i are 1 and 0.5. Given a mail with specific w_i, $(-2)\ln(p_1\times p_2\times ...\times p_n)$ will follow a χ^2 distribution with degree of freedom in $2n$. We can simply use a inverse χ^2 distribution function to derive the probability of the mail that being spam.

$$H=C^{-1}[-2\ln\prod_{w} f(w), 2n] \tag{5}$$

ID3 Decision Tree

The ID3.decision tree is based on information theory and attempts to minimize the expected number of comparisons. The basic strategy is to choose splitting attributes with the highest information gain (or highest entropy reduction) first. Such an approach will minimize the expected number of tests needed to

classify an object and guarantees that a simple (but not necessarily the simplest) tree is found. Given probabilities $p_1, p_2, ..., p_n$, where $\Sigma_{i=1}^{n} p_i = 1$, entropy is defined as:

$$I(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (6)$$

See [2] for the detailed procedure.

EXPERIMENTS AND RESULT

We use the Spam Email Database from the UCI Machine Learning Repository for the experiment. We randomly choose 50% instances (2282) for algorithm training, and 2319 instances for testing. There are 908 spam and 1411 legitimate mail in the testing data. To evaluate the filtering performance, we employ several quantities typically used in measuring the query result in information retrieval. The result is shown in Table 1:

Table 1: Results of the experiments

	True Positives	True Negatives	False Positives	False Negatives	Precision	Recall	Accuracy
Naive Bayes	719	1299	46	255	94%	74%	87%
Fisher's Method	835	1229	116	139	89%	87%	90%
Decision Tree	881	1275	70	93	92%	91%	93%

As we can see, Naïve Bayes has the highest precision rate, but the recall and accuracy rates are not as good as others. And it suffers from the false negatives rate. The Fisher's method has better recall and accuracy rates than Naïve Bayes, though the precision rate is the lowest of the three. The decision tree method generally has better performance than others.

CONCLUSION

We examined three data mining methods for spam detection. From the result, it is possible to build a filtering model through data mining algorithms. But currently the methods require a significant amount of memory and computing resource. In our future work, we will develop more efficient algorithms and make the learning algorithms more efficient in time and space.

Acknowledgement: The author would like to give thanks to the National Science Council of Taiwan for grant (NSC93-2213-E-224-038) to part of this research.