A Three-Way Decision Approach to Email Spam Filtering*

Bing Zhou, Yiyu Yao, and Jigang Luo

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{zhou200b,yyao,luo226}@cs.uregina.ca

Abstract. Many classification techniques used for identifying spam emails, treat spam filtering as a binary classification problem. That is, the incoming email is either spam or non-spam. This treatment is more for mathematical simplicity other than reflecting the true state of nature. In this paper, we introduce a three-way decision approach to spam filtering based on Bayesian decision theory, which provides a more sensible feedback to users for precautionary handling their incoming emails, thereby reduces the chances of misclassification. The main advantage of our approach is that it allows the possibility of rejection, i.e., of refusing to make a decision. The undecided cases must be re-examined by collecting additional information. A loss function is defined to state how costly each action is, a pair of threshold values on the posterior odds ratio is systematically calculated based on the loss function, and the final decision is to select the action for which the overall cost is minimum. Our experimental results show that the new approach reduces the error rate of classifying a legitimate email to spam, and provides better spam precision and weighted accuracy.

Key words: spam filter, three-way decision, naive Bayesian classification, Bayesian decision theory, cost

1 Introduction

Email spam filtering is a growing concern on the Internet. A popular approach is to treat spam filtering as a classification problem. Many classification algorithms from machine learning were employed to automatically classify incoming emails into different categories based on the contents of emails [2, 6, 9, 11, 14, 15]. Among these algorithms, Bayesian classifier achieved better results by reducing the classification error rates. The naive Bayesian classifier [6, 11, 14], along with many other classification algorithms, treat spam filtering as a binary classification problem, that is, the incoming email is either spam or non-spam. In reality, this simple treatment is too restrict and could result in losing vital information

* Bing Zhou, Yiyu Yao and Jigang Luo, A Three-Way Decision Approach to Email Spam Filtering, Proceedings of the 23rd Canadian Conference on Artificial Intelligence, LNAI 6085, pp. 28-39, 2010.

for users by misclassifying a legitimate email to spam. For example, a user could miss an important job offer just because the email contains "congratul" (i.e., a common word in email spam filter word list) in its header. On the other hand, misclassifying a spam email to non-spam also brings unnecessary costs and waste of resources.

In this paper, we introduce a three-way decision approach to spam filtering based on Bayesian decision theory, that is, to *accept*, *reject*, or *further-exam* an incoming email. The emails waiting for *further-exam* must be clarified by collecting additional information. The idea of three-way decision making can be found in some early literatures and has been applied to many real world problems [5, 7]. For example, the three-way decisions are often used in clinical decision making for a certain disease, with options of treating the conditional directly, not treating the condition, or performing a diagnose test to decide whether or not to treat the condition [12]. Yao et al. [16, 17] introduced decision theoretic rough set model (DTRS) based on three-way decisions. The ideas of DTRS have been applied to information retrieval by dividing the dynamic document stream into three states instead of the traditional relevant and irrelevant states [8]. More recently, Zhao et al. [18] introduced an email classification schema based on DTRS by classifying the incoming email into three categories instead of two. The main differences between their work and our approach are the interpretations of the conditional probabilities and the values of the loss functions. In their approach, the conditional probability was estimated by the rough membership function [13], which is only one of the possible ways and is impractical for real applications. They have simply defined the loss function that all errors are treated equally, which is not the case in many real applications. For instance, misclassifying a legitimate email to spam is usually considered more costly than misclassifying a spam email to legitimate. In our approach, the conditional probability is interpreted based on the naive Bayesian classification. The posterior odds is used a monotonic increasing transformation of the conditional probability to compare with the threshold values. A threshold value on the probability can indeed be interpreted as another threshold value on the odds. The naive independence assumptions are added to calculate the likelihood by assuming that each feature of an email is unrelated to any other features. After the transformations, all the related factors used to interpret the conditional probability are easily derivable from data. We consider the different cost associated for taking each action, which is more general than the zero-one loss function.

The main advantage of three-way decision making is that it allows the possibility of rejection, i.e., of refusing to make a decision. The undecided cases must be forwarded for re-examination. A loss function is defined to state how costly each action is, and the final decision is to select the action for which the overall cost is minimum. A pair of threshold values are estimated based on the loss function. The first threshold value determines the value of the probability necessary for a re-examination, and the second value determines the value of the probability necessary to reject an email. These settings provide users a fairly high degree of control over their incoming emails, thereby reduce the chances of

misclassification. Our experimental results show that the new approach reduces the error rate of classifying a legitimate email to spam, and provides a better spam precision and weighted accuracy.

2 The Naive Bayesian Spam Filtering

The naive Bayesian spam filtering is a probabilistic classification technique of email filtering [14]. It is based on Bayes' theorem with naive (strong) independence assumptions [6, 11, 14].

Suppose each email can be described by a feature vector $\mathbf{x} = (x_1, x_2, ..., x_n)$, where $x_1, x_2, ..., x_n$ are the values of attributes of emails. Let C denote the *legitimate* class, and C^c denote the *spam* class. Based on Bayes' theorem and the theorem of total probability, given the vector of an email, the conditional probability that this email is in the *legitimate* class is:

$$Pr(C|\mathbf{x}) = \frac{Pr(C)Pr(\mathbf{x}|C)}{Pr(\mathbf{x})}, \tag{1}$$

where $Pr(\mathbf{x}) = Pr(\mathbf{x}|C)Pr(C) + Pr(\mathbf{x}|C^c)Pr(C^c)$. Here $Pr(C)$ is the *prior* probability of an email being in the *legitimate* class. $Pr(\mathbf{x}|C)$ is commonly known as the *likelihood* of an email being in the *legitimate* class with respect to $\mathbf{x}$.

The likelihood $Pr(\mathbf{x}|C)$ is a joint probability of $Pr(x_1, x_2, ..., x_n|C)$. In practice, it is difficult to analyze the interactions between the components of $\mathbf{x}$, especially when the number n is large. In order to solve this problem, an independence assumption is embodied in the naive Bayesian classifier [6, 11] which assumes that each feature x_i is conditionally independent of every other features, given the class C, this yields,

$$\begin{aligned} Pr(\mathbf{x}|C) &= Pr(x_1, x_2, ..., x_n|C) \\ &= \prod_{i=1}^{n} Pr(x_i|C), \end{aligned} \tag{2}$$

where $Pr(x_i|C)$ can be easily estimated as relative frequencies from the training data set. Thus equation (1) can be rewritten as:

$$Pr(C|\mathbf{x}) = \frac{Pr(C)\prod_{i=1}^{n} Pr(x_i|C)}{Pr(\mathbf{x})}. \tag{3}$$

Similarly, the corresponding probabilities $Pr(C^c|\mathbf{x})$ of an email being in the *spam* class given vector $\mathbf{x}$ can be reformulated as:

$$Pr(C^c|\mathbf{x}) = \frac{Pr(C^c)\prod_{i=1}^{n} Pr(x_i|C^c)}{Pr(\mathbf{x})}. \tag{4}$$

Note that $Pr(\mathbf{x})$ in equation (3) and (4) is unimportant with regard to making a decision. It is basically a scale factor that assures $Pr(C|\mathbf{x}) + Pr(C^c|\mathbf{x}) = 1$. This

scale factor can be eliminated by taking the ratio of $Pr(C|\mathbf{x})$ and $Pr(C^c|\mathbf{x})$:

$$\frac{Pr(C|\mathbf{x})}{Pr(C^c|\mathbf{x})} = \prod_{i=1}^{n} \frac{Pr(x_i|C)}{Pr(x_i|C^c)} \frac{Pr(C)}{Pr(C^c)}. \tag{5}$$

$Pr(C|\mathbf{x})/Pr(C^c|\mathbf{x})$ is called the *posterior odds* of an email being in the *legitimate* class against being in the *spam* class given $\mathbf{x}$. It is a monotonic increasing transformation of $Pr(C|\mathbf{x})$. $Pr(x_i|C)/Pr(x_i|C^c)$ is called the *likelihood ratio*. Thus, the conditional probability $Pr(C|\mathbf{x})$ can be easily calculated from $Pr(C|\mathbf{x})/Pr(C^c|\mathbf{x})$ based on the observation that $Pr(C|\mathbf{x}) + Pr(C^c|\mathbf{x}) = 1$. Finally, an incoming email can be classified as *legitimate* if $\frac{Pr(C|\mathbf{x})}{Pr(C^c|\mathbf{x})}$ (i.e., the posterior odds) exceeds a threshold value, otherwise it is *spam*.

3 Bayesian Decision Theory

Bayesian decision theory is a fundamental statistical approach that makes decisions under uncertainty based on probabilities and costs associated with decisions. Following the discussions given in the book by Duda and Hart [3], the basic ideas of the theory are reviewed.

Let $\Omega = \{w_1, \ldots, w_s\}$ be a finite set of s states and let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be a finite set of m possible actions. Let $\lambda(a_i|w_j)$ denote the loss, or cost, for taking action a_i when the state is w_j. Let $Pr(w_j|\mathbf{x})$ be the conditional probability of an email being in state w_j given that the email is described by $\mathbf{x}$. For an email with description $\mathbf{x}$, suppose action a_i is taken. Since $Pr(w_j|\mathbf{x})$ is the probability that the true state is w_j given $\mathbf{x}$, the expected loss associated with taking action a_i is given by:

$$R(a_i|\mathbf{x}) = \sum_{j=1}^{s} \lambda(a_i|w_j) Pr(w_j|\mathbf{x}). \tag{6}$$

The quantity $R(a_i|\mathbf{x})$ is also called the conditional risk.

Given a description $\mathbf{x}$, a decision rule is a function $\tau(\mathbf{x})$ that specifies which action to take. That is, for every $\mathbf{x}$, $\tau(\mathbf{x})$ takes one of the actions, $a_1, \ldots, a_m$. The overall risk $\mathbf{R}$ is the expected loss associated with a given decision rule. Since $R(\tau(\mathbf{x})|\mathbf{x})$ is the conditional risk associated with action $\tau(\mathbf{x})$, the overall risk is defined by:

$$\mathbf{R} = \sum_{\mathbf{x}} R(\tau(\mathbf{x})|\mathbf{x}) Pr(\mathbf{x}), \tag{7}$$

where the summation is over the set of all possible descriptions of emails. If $\tau(\mathbf{x})$ is chosen so that $R(\tau(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$, the overall risk $\mathbf{R}$ is minimized. Thus, the optimal Bayesian decision procedure can be formally stated as follows. For every $\mathbf{x}$, compute the conditional risk $R(a_i|\mathbf{x})$ for $i = 1, \ldots, m$ defined by equation (6) and select the action for which the conditional risk is minimum. If more than one action minimizes $R(a_i|\mathbf{x})$, a tie-breaking criterion can be used.

4 A Three-Way Decision Approach to Email Spam Filtering

In the naive Bayesian spam filter, an incoming email is classified as legitimate if the posterior odds ratio exceeds a certain threshold value. In our approach, a pair of threshold values is used to make a three-way decision of an incoming email. The first threshold value determines the probability necessary for a re-examination, and the second value determines the probability necessary to reject an email. There are different ways to acquire the required threshold values. One may directly supply the threshold values based on an intuitive understanding of the levels of tolerance for errors [19]. A more rational way is to infer these threshold values from a theoretical and practical basis. One such solution was given in DTRS [16, 17] based on the well known Bayesian decision theory [3]. A pair of threshold values on the conditional probability is systematically calculated based on the loss function. In our approach, the posterior odds is used a monotonic increasing transformation of the conditional probability to compare with the threshold values. A new pair of threshold values is defined and calculated based on the prior odds ratio and the loss functions with the naive independence assumptions. This transformation ensures the easy estimation of all the related factors.

With respect to a set of emails to be approximated, we have a set of two states $\Omega = \{C, C^c\}$ indicating that an email is in C (i.e., *legitimate*) or not in C (i.e., *spam*), respectively. The incoming emails can be divided into three regions, namely, the positive region $\mathrm{POS}(C)$ includes emails being *legitimate*, the boundary region $\mathrm{BND}(C)$ includes emails that need *further-exam*, and the negative region $\mathrm{NEG}(C)$ includes emails that are *spam*. With respect to these three regions, the set of actions is given by $\mathcal{A} = \{a_P, a_B, a_N\}$, where a_P, a_B, and a_N represent the three actions in classifying an email x, namely, deciding $x \in \mathrm{POS}(C)$, deciding $x \in \mathrm{BND}(C)$, and deciding $x \in \mathrm{NEG}(C)$, respectively. The loss function is given by the 3×2 matrix:

	C (P)	C^c (N)
a_P	$\lambda_{PP} = \lambda(a_P\|C)$	$\lambda_{PN} = \lambda(a_P\|C^c)$
a_B	$\lambda_{BP} = \lambda(a_B\|C)$	$\lambda_{BN} = \lambda(a_B\|C^c)$
a_N	$\lambda_{NP} = \lambda(a_N\|C)$	$\lambda_{NN} = \lambda(a_N\|C^c)$

In the matrix, λ_{PP}, λ_{BP} and λ_{NP} denote the losses incurred for taking actions a_P, a_B and a_N, respectively, when an email belongs to C, and λ_{PN}, λ_{BN} and λ_{NN} denote the losses incurred for taking these actions when the email does not belong to C.

The expected losses associated with taking different actions for emails with description $\mathbf{x}$ can be expressed as:

$$\begin{aligned} R(a_P|\mathbf{x}) &= \lambda_{PP} Pr(C|\mathbf{x}) + \lambda_{PN} Pr(C^c|\mathbf{x}), \\ R(a_B|\mathbf{x}) &= \lambda_{BP} Pr(C|\mathbf{x}) + \lambda_{BN} Pr(C^c|\mathbf{x}), \\ R(a_N|\mathbf{x}) &= \lambda_{NP} Pr(C|\mathbf{x}) + \lambda_{NN} Pr(C^c|\mathbf{x}). \end{aligned} \tag{8}$$

The Bayesian decision procedure suggests the following minimum-risk decision rules:

(P) If $R(a_P|\mathbf{x}) \leq R(a_B|\mathbf{x})$ and $R(a_P|\mathbf{x}) \leq R(a_N|\mathbf{x})$, decide $x \in \text{POS}(C)$;
(B) If $R(a_B|\mathbf{x}) \leq R(a_P|\mathbf{x})$ and $R(a_B|\mathbf{x}) \leq R(a_N|\mathbf{x})$, decide $x \in \text{BND}(C)$;
(N) If $R(a_N|\mathbf{x}) \leq R(a_P|\mathbf{x})$ and $R(a_N|\mathbf{x}) \leq R(a_B|\mathbf{x})$, decide $x \in \text{NEG}(C)$.

Tie-breaking criteria should be added so that each email is put into only one region.

Since $Pr(C|\mathbf{x}) + Pr(C^c|\mathbf{x}) = 1$, we can simplify the rules based only on the probabilities $Pr(C|\mathbf{x})$ and the loss function λ. Consider a special kind of loss functions with:

$$\begin{aligned} \text{(c0).} \quad & \lambda_{PP} \leq \lambda_{BP} < \lambda_{NP}, \\ & \lambda_{NN} \leq \lambda_{BN} < \lambda_{PN}. \end{aligned} \tag{9}$$

That is, the loss of classifying an email x being in C into the positive region $\text{POS}(C)$ is less than or equal to the loss of classifying x into the boundary region $\text{BND}(C)$, and both of these losses are strictly less than the loss of classifying x into the negative region $\text{NEG}(C)$. The reverse order of losses is used for classifying an email not in C. Under condition (c0), we can simplify decision rules (P)-(N) as follows. For the rule (P), the first condition can be expressed as:

$$\begin{aligned} & R(a_P|\mathbf{x}) \leq R(a_B|\mathbf{x}) \\ \Longleftrightarrow\ & \lambda_{PP}Pr(C|\mathbf{x}) + \lambda_{PN}Pr(C^c|\mathbf{x}) \leq \lambda_{BP}Pr(C|\mathbf{x}) + \lambda_{BN}Pr(C^c|\mathbf{x}) \\ \Longleftrightarrow\ & \lambda_{PP}Pr(C|\mathbf{x}) + \lambda_{PN}(1 - Pr(C|\mathbf{x})) \leq \lambda_{BP}Pr(C|\mathbf{x}) + \lambda_{BN}(1 - Pr(C|\mathbf{x})) \\ \Longleftrightarrow\ & Pr(C|\mathbf{x}) \geq \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}. \end{aligned} \tag{10}$$

Similarly, the second condition of rule (P) can be expressed as:

$$R(a_P|\mathbf{x}) \leq R(a_N|\mathbf{x}) \Longleftrightarrow Pr(C|\mathbf{x}) \geq \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}. \tag{11}$$

The first condition of rule (B) is the converse of the first condition of rule (P). It follows,

$$R(a_B|\mathbf{x}) \leq R(a_P|\mathbf{x}) \Longleftrightarrow Pr(C|\mathbf{x}) \leq \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}. \tag{12}$$

For the second condition of rule (B), we have:

$$R(a_B|\mathbf{x}) \leq R(a_N|\mathbf{x}) \Longleftrightarrow Pr(C|\mathbf{x}) \geq \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}. \tag{13}$$

The first condition of rule (N) is the converse of the second condition of rule (P) and the second condition of rule (N) is the converse of the second condition of

rule (B). It follows,

$$R(a_N|\mathbf{x}) \leq R(a_P|\mathbf{x}) \Longleftrightarrow Pr(C|\mathbf{x}) \leq \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})},$$
$$R(a_N|\mathbf{x}) \leq R(a_B|\mathbf{x}) \Longleftrightarrow Pr(C|\mathbf{x}) \leq \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}. \quad (14)$$

To obtain a compact form of the decision rules, we denote the three expressions in these conditions by the following three parameters:

$$\alpha = \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})},$$
$$\beta = \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})},$$
$$\gamma = \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}. \quad (15)$$

The decision rules (P)-(N) can be expressed concisely as:

(P) If $Pr(C|\mathbf{x}) \geq \alpha$ and $Pr(C|\mathbf{x}) \geq \gamma$, decide $x \in \text{POS}(C)$;
(B) If $Pr(C|\mathbf{x}) \leq \alpha$ and $Pr(C|\mathbf{x}) \geq \beta$, decide $x \in \text{BND}(C)$;
(N) If $Pr(C|\mathbf{x}) \leq \beta$ and $Pr(C|\mathbf{x}) \leq \gamma$, decide $x \in \text{NEG}(C)$.

Each rule is defined by two out of the three parameters.

The conditions of rule (B) suggest that $\alpha > \beta$ may be a reasonable constraint; it will ensure a well-defined boundary region. By setting $\alpha > \beta$, namely,

$$\frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})} > \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}, \quad (16)$$

we obtain the following condition on the loss function [17]:

$$\text{(c1).} \qquad \frac{\lambda_{NP} - \lambda_{BP}}{\lambda_{BN} - \lambda_{NN}} > \frac{\lambda_{BP} - \lambda_{PP}}{\lambda_{PN} - \lambda_{BN}}. \quad (17)$$

The condition (c1) implies that $1 \geq \alpha > \gamma > \beta \geq 0$. In this case, after tie-breaking, the following simplified rules are obtained [17]:

(P1) If $Pr(C|\mathbf{x}) \geq \alpha$, decide $x \in \text{POS}(C)$;
(B1) If $\beta < Pr(C|\mathbf{x}) < \alpha$, decide $x \in \text{BND}(C)$;
(N1) If $Pr(C|\mathbf{x}) \leq \beta$, decide $x \in \text{NEG}(C)$.

The parameter γ is no longer needed.

¿From the rules (P1), (B1), and (N1), the (α, β)-probabilistic positive, negative and boundary regions are given, respectively, by:

$$\text{POS}_{(\alpha,\beta)}(C) = \{x \in U \mid Pr(C|\mathbf{x}) \geq \alpha\},$$
$$\text{BND}_{(\alpha,\beta)}(C) = \{x \in U \mid \beta < Pr(C|\mathbf{x}) < \alpha\},$$
$$\text{NEG}_{(\alpha,\beta)}(C) = \{x \in U \mid Pr(C|\mathbf{x}) \leq \beta\}. \quad (18)$$

The threshold parameters can be systematically calculated from a loss function based on the Bayesian decision theory.

The conditional probability $Pr(C|\mathbf{x})$ is difficult to directly derive from data. Recall that in naive Bayesian spam filter, the ratio of $Pr(C|\mathbf{x})$ and $Pr(C^c|\mathbf{x})$ (i.e., the posterior odds) can be used as a monotonic increasing transformation of the conditional probability $Pr(C|\mathbf{x})$. A threshold value on the probability can indeed be interpreted as another threshold value on the odds. For the positive region, we have:

$$P(C|\mathbf{x}) \geq \alpha \Longleftrightarrow \frac{Pr(C|\mathbf{x})}{Pr(C^c|\mathbf{x})} \geq \frac{\alpha}{1-\alpha} = \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{BP} - \lambda_{PP}}. \tag{19}$$

According to equation (5), we can re-expressed the above equation as:

$$\prod_{i=1}^{n} \frac{Pr(x_i|C)}{Pr(x_i|C^c)} \frac{Pr(C)}{Pr(C^c)} \geq \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{BP} - \lambda_{PP}}. \tag{20}$$

This computation can be further simplified by taking the logarithm of both side of equation (20):

$$\sum_{i=1}^{n} \log \frac{Pr(x_i|C)}{Pr(x_i|C^c)} \geq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{BP} - \lambda_{PP}}. \tag{21}$$

Here $\log \frac{Pr(C^c)}{Pr(C)}$ is independent of the description of emails, we treat it as a constant. Similar expression can be obtained for the negative region as:

$$\sum_{i=1}^{n} \log \frac{Pr(x_i|C)}{Pr(x_i|C^c)} \leq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\lambda_{BN} - \lambda_{NN}}{\lambda_{NP} - \lambda_{BP}}. \tag{22}$$

A new pair of threshold values α' and β' can be defined as:

$$\begin{aligned} \alpha' &= \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{BP} - \lambda_{PP}}, \\ \beta' &= \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\lambda_{BN} - \lambda_{NN}}{\lambda_{NP} - \lambda_{BP}}, \end{aligned} \tag{23}$$

where $Pr(C)/Pr(C^c)$ can be easily estimated from the frequencies of the training data by putting:

$$Pr(C) = \frac{|C|}{|U|} \quad \text{and} \quad Pr(C^c) = \frac{|C^c|}{|U|}. \tag{24}$$

Table 1. Three-way decision results with $\lambda = 1$

	Actually legitimate	Actually spam	Total
accept	465	28	493
further-exam	22	13	35
Reject	12	227	239
Total	499	268	767

Table 2. Naive Bayesian results with $\lambda = 1$

	Actually legitimate	Actually spam	Total
Classified legitimate	476	32	508
Classified spam	23	236	259
Total	499	268	767

We can then get the (α', β')-probabilistic positive, negative and boundary regions written as:

$$\begin{aligned}
\mathrm{POS}_{(\alpha',\beta')}(C) &= \{x \in U \mid \sum_{i=1}^{n} \log \frac{Pr(x_i|C)}{Pr(x_i|C^c)} \geq \alpha'\}, \\
\mathrm{BND}_{(\alpha',\beta')}(C) &= \{x \in U \mid \beta' < \sum_{i=1}^{n} \log \frac{Pr(x_i|C)}{Pr(x_i|C^c)} < \alpha'\}, \\
\mathrm{NEG}_{(\alpha',\beta')}(C) &= \{x \in U \mid \sum_{i=1}^{n} \log \frac{Pr(x_i|C)}{Pr(x_i|C^c)} \leq \beta'\}.
\end{aligned} \tag{25}$$

All the factors in equation (25) are easy to derive from data.

5 Experimental Results and Evaluations

Our experiments were performed on a spambase data set from UCI Machine Learning Repository [10]. The data set consists of 4601 instances, with 1813 instances as *spam*, and 2788 instances as *legitimate*, each instance is described by 58 attributes. Our goal is to compare our approach with the original naive Bayesian spam filter in terms of the error rate that a legitimate email is classified as spam, the precision and recall for both legitimate and spam emails, and the cost-sensitive measure suggested by Androutsopoulos et al. [1].

We split the spambase data set into a training set of 3834 instances, and a testing set of 767 instances. Since the attributes in the input data set have continuous values, entropy-MDL [4] is used as the discretization method applied to both the training and testing data sets before the calculations of probabilities. For the cost-sensitive evaluations, we assume that misclassifying a legitimate email as spam is λ times more costly than misclassifying a spam email as legitimate. We considered three different λ values ($\lambda = 9$, $\lambda = 3$, and $\lambda = 1$) for the original naive Bayesian spam filter. Three sets of loss functions for the three-way decision approach are set up accordingly with the same cost ratios. For instance,

Table 3. Three-way decision results with $\lambda = 3$

	Actually legitimate	Actually spam	Total
accept	476	32	508
further-exam	12	10	22
Reject	11	226	237
Total	499	268	767

Table 4. Naive Bayesian results with $\lambda = 3$

	Actually legitimate	Actually spam	Total
Classified legitimate	483	38	521
Classified spam	16	230	246
Total	499	268	767

Table 5. Three-way decision results with $\lambda = 9$

	Actually legitimate	Actually spam	Total
accept	465	28	493
further-exam	29	36	65
Reject	5	204	209
Total	499	268	767

Table 6. Naive Bayesian results with $\lambda = 9$

	Actually legitimate	Actually spam	Total
Classified legitimate	491	46	537
Classified spam	8	222	230
Total	499	268	767

when we use $\lambda = 9$ for the naive Bayesian spam filter, $\lambda_{NP}/\lambda_{PN} = 9$ is used in the three-way decision approach.

Table 1 and Table 2 show the prediction results of the three-way decision and the naive Bayesian approach when $\lambda = 1$, respectively. Note that in this case, the cost of misclassifying a legitimate email as spam is the same as the cost of misclassifying a spam email as legitimate. Table 3 and Table 4 show the prediction results when $\lambda = 3$. Table 5 and Table 6 show the prediction results when $\lambda = 9$. From the above tables, we can easily find that the error rates of misclassifying a legitimate email into spam by using the three-way decision approach are lower than the original naive Bayesian spam filter in all three experiments. Since reducing this error rate is the most important factor to users. Although the accuracy of correctly classifying a legitimate email has slightly dropped, but we consider this as a reasonable trade off.

To further evaluate these results, we compare the precision, recall and weighted accuracy of both approaches. The legitimate precision and recall are defined as:

$$\textit{legitimate precision} = \frac{n_{L\to L}}{n_{L\to L} + n_{S\to L}}, \quad \textit{legitimate recall} = \frac{n_{L\to L}}{n_{L\to L} + n_{L\to S}},$$

where $n_{L\to L}$ denotes the number of emails classified as legitimate which truly are, $n_{L\to S}$ denotes the number of legitimate emails classified as spam, and $n_{S\to L}$

Table 7. Comparison between three-way decision and naive Bayesian approaches

Cost	Approaches	Spam		Legitimate		
		precision	recall	precision	recall	weighted accuracy
$\lambda = 1$	Three-way decision	94.98%	84.70%	94.32%	93.19%	94.54%
	Naive Bayesian	91.12%	88.06%	93.70%	95.39%	92.83%
$\lambda = 3$	Three-way decision	95.36%	84.33%	93.70%	90.71%	96.22%
	Naive Bayesian	93.50%	85.82%	92.70%	96.79%	95.13%
$\lambda = 9$	Three-way decision	97.61%	76.12%	94.32%	93.19%	98.36%
	Naive Bayesian	96.52%	82.84%	91.43%	98.40%	97.52%

denotes the number of spam emails classified as legitimate. Similarly, we define:

$$\textit{spam precision} = \frac{n_{S\to S}}{n_{S\to S} + n_{L\to S}}, \quad \textit{spam recall} = \frac{n_{S\to S}}{n_{S\to S} + n_{S\to L}}.$$

Clearly, spam precision is the most important factor to users. The comparison results are shown in Table 7. We can easily find that the three-way decision approach provides a better spam precision than the naive Bayesian spam filter in all three experiments. For the cost-sensitive evaluations, we used weighted accuracy suggested by Androutsopoulos et al. [1], which is defined as:

$$\textit{weighted accuracy} = \frac{\lambda \cdot n_{L\to L} + n_{S\to S}}{\lambda \cdot N_L + N_S},$$

where N_L and N_S are the number of legitimate and spam emails to be classified by the spam filter. From Table 7, we can find that the weighted accuracy of the three-way decision approach is higher than the original naive Bayesian approach in all three experiments. We also find that when λ changed to a bigger value, the performances of both approaches are increased, but the three-way decision approach performs out the naive Baysian spam filter in all three settings.

6 Conclusion

In this paper, we present a three-way decision approach to email spam filtering. In addition to the most commonly used binary classification for spam filtering, a third action is added to allow users make further examinations for undecided cases. The main advantage of our approach is that it provides a more sensible feedback to users for handling their emails, thus reduces the misclassification rate. A pair of threshold values are used. The first threshold value determines the point necessary for a re-examination, and the second value determines the point to reject an email. Instead of supplying the threshold values based on try and error, or intuitive understandings of the levels of tolerance for errors. We provide a systematically calculation of the threshold values based on Bayesian decision theory. A loss function is defined in association with each action. The final decision making is to select the action for which the overall cost is minimum.

Our experimental results show that the new approach reduces the error rate of classifying a legitimate email to spam, and provides a better spam precision and weighted accuracy.

Acknowledgements

The first author is supported by an NSERC Alexander Graham Bell Canada Graduate Scholarship. The second author is partially supported by an NSERC Canada Discovery grant.

References

1. Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., and Spyropoulos, C.D. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 160-167, 2000.
2. Cristianini, N., and Shawe-Taylor, J. *An Introduction to Support vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.
3. Duda, R.O., and Hart, P.E. *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
4. Fayyad, U.M., Irani, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022-1029, 1993.
5. Forster, M.R. Key concepts in model selection: performance and generalizability, *Journal of Mathematical Psychology*, **44**, pp. 205-231, 2000.
6. Good, I.J., *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. M.I.T. Press. 1965.
7. Goudey, R. Do statistical inferences allowing three alternative decision give better feedback for environmentally precautionary decision-making, *Journal of Environmental Management*, **85**, pp. 338-344, 2007.
8. Li, Y.F., and Zhang, C.Q. Rough set based decision model in information retrieval and filtering, *Third World Multiconference on Systemics, Cybernetics and Informatics (SCI'99) and Fifth International Conference on Information Systems Analysis and Synthesis (ISAS'99)*, **5**, pp. 398-403, 1999.
9. Masand, B., Linoff, G., and Waltz D. Classifying news stories using memory based reasoning. *In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 59-65, 1992.
10. http://www.ics.uci.edu/mlearn/MLRepository.html.
11. Mitchell, T., *Machine Learning*. McGraw Hill, 1997.
12. Pauker, S.G., and Kassirer, J.P. The threshold approach to clinical decision making, *New England Journal of Medicine*, 1980.
13. Pawlak, Z., Skowron, A. Rough membership functions, in: Yager, R.R., Fedrizzi, M. and Kacprzyk, J., Eds., *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley and Sons, New York, pp. 251-271, 1994.
14. Sahami, M., Dumais, S., Heckerman. D., and Horvitz, E., A Bayesian approach to filtering junk e-mail, *AAAI Workshop on Learning for Text Categorization*, Madison, Wisconsin. AAAI Technical Report WS-98-05, 1998.

15. Schapire, E., and Singer Y. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, **39(2/3)**, pp. 135-168, 2000.
16. Yao, Y.Y., Wong, S.K.M., and Lingras, P. A decision-theoretic rough set model, in: *Methodologies for Intelligent Systems 5*, Z.W. Ras, M. Zemankova and M.L. Emrich (Eds.), New York, North-Holland, pp. 17-24, 1990.
17. Yao, Y.Y. Decision-theoretic rough set models, *Proceedings of RSKT 2007*, LNAI 4481, pp. 1-12, 2007.
18. Zhao, W.Q., Zhu, Y.L. An email classification scheme based on decision-theoretic rough set theory and analysis of email security, *Proceeding of 2005 IEEE Region 10 TENCON*, pp. 1-6, 2005.
19. Ziarko, W. Variable precision rough sets model, *Journal of Computer and Systems Sciences*, **46**, pp. 39C59, 1993.