Application of Boosting Algorithm in Spam Filtration

Gu Wencheng
Network and Information Center, Qiqihar University
Qiqihar, Heilongjiang, China, 161006
email: guwencheng@163.com

Abstract

In order to improve accuracy and effectiveness of junk mail filtration, a filtering method is proposed based on Boosting algorithm, which uses Boosting algorithm to construct a spam filterer to identify junk mail. Besides, reference technical indexes in the field of text classification and information retrieval are used to construct a spam filterer evaluation system, with it, experimental data obtained from simulation were tested and evaluated. The results of test and evaluation proved that, compared with the traditional Bayesian algorithm, the spam filterer based on Boosting algorithm is able to filter junk mails more excellently, and the effectiveness of Boosting algorithm in spam filtering is verified.

Keywords: boosting algorithm, junk mail, filtration, classifier, evaluation

1. Introduction


In theory, Boosting algorithm is an algorithm framework that can integrate any weak classification algorithms, and has a relatively complete basis of mathematical theory and experiments show that, Boosting algorithm has strong applicability for a small size of samples and high-dimensional data, compared with other classification algorithms. Boosting algorithm is fast, simple, easily programmed with high adaptability and accuracy, and capable of feature selection while classifying.

With the increasing development and improvement of Boosting algorithm, it has been widely used in more and more areas. In the field of image recognition and retrieval, Boosting algorithm has been successfully applied in handwritten character recognition [5], and OCR character recognition [6]; in terms of speed and accuracy of face detection, Boosting algorithm has achieved good results [7]; in the field of medical diagnosis, it has been used in lung and skin cancer diagnosis [8]; as for biological information processing, it has been employed in gene expression profiles classification [9]; in military field, it has successfully been applied to identify radar signals [10]; in the field of health care, sub-health groups are classified by Boosting algorithm [11]; in addition, Boosting algorithm has also been applied in intrusion detection technology [12], semi-structured information extraction [13], target tracking [14], oil flooded layer identification [15], human action recognition [16] and other fields. Here, with the help of Boosting algorithm, we filter junk mail.

This paper is organized as follows. The theoretical basis is presented for this work in Section 2. In Section 3, Boosting algorithm is designed for junk mail filtering. In Section 4, evaluation system of junk mail filtering based on Boosting algorithm is constructed. In Section 5, detailed information on the experimental results and analysis is discussed and summarized. In Section 6, conclusions are drawn.
2. Boosting Algorithm

Boosting algorithm can raise the recognition rate of weak classification algorithm, whose core idea is, viewing the other weak classification algorithm as base classification algorithms, to put them into the framework of Boosting algorithms, under which the sample set training is operated, so that different training samples subsets are obtained, and then by training these samples subsets base classifiers are obtained; after \( N \) round of such given training, \( N \) base classifiers are generated; Boosting algorithms ,by weighted fusion of these \( N \) base classifiers, produce a final classifier . While for the \( N \) base classifiers, each individual classifier recognition rate is not very high, after the weighted fusion the final classifier often has a higher recognition rate, so as to improve the weak classification algorithm the recognition ratio.

Boosting algorithm is an iterative algorithm, whose specific operation is to construct a series of predictive function, then in a certain way they are combined into a predictive function. The basic idea of the Boosting algorithm is : in a given weak learning algorithm and a total sample set \( U: (u_1,v_1),(u_2,v_2),..., (u_n,v_n) \), \( u_i \) is the i-th training samples input, \( v_i \in V = \{+1,-1\} \) is the class mark of classification problems ; in Boosting algorithm, first the training sample weight distribution is initialized with \( 1/n \) as the specified training set distribution, that is, the sample weight \( D_i \) for each training \( U_i \) is \( 1/n \), and then the appropriate weak learning algorithm is used for \( N \) iterations , after each iteration , according to the results of the training the distribution of the training set is updated, for those failed training samples the weight is redistributed to a greater one , so that in the next iteration, more attention is to be paid on these training samples; at the end of the iteration , there is a prediction function \( H(U) \) sequence \( h_1,h_2,...,h_n \), where each prediction function \( h_i \) is corresponded with a weight value \( D_i \), for the prediction function with excellent performance, the corresponding weight value is greater, whereas the prediction function with bad performance, the corresponding weight value is smaller. After \( N \) iterations, the final prediction function \( H(U) \) is generated by a joint weighted fusion \( w_i \). For a single weak learning algorithm, its learning accuracy rate is not high, but after Boosting algorithm, the accuracy of the final results will be greatly improved. Algorithm in Figure 1 can be used to describe the process.

3. Filtering of Junk Mail Based on Boosting Algorithm

E-mail is essential to people's work and life as the information transmission means, bringing them convenience, but at the same time it also causes a lot of new problems, the main problem is the emergence of a large number of junk mail, so junk mail filtering has become one of the issues the majority of e-mail users are concerned about most.

Currently, junk mail filtering methods consist of filtering based on IP address, filtering based on behavioral features, and filtering based on message and so on. The IP address filtering techniques is to restrict or filter by e-mail address, IP or "black/white list" of domain name [17]; the behavioral features filtering is to determine whether the message is junk by behavior features of junk mail different from the normal mail, such as special time to send, the
high transmission frequency in a short term, mail forwarding, abnormal e-mail address, abnormal SMTP session information, attacking other hosts, several transit routers, false server information, e-mail header information abnormalities, and hidden sending address; content filtering technique [18] is to take e-mail message header, sender, recipient, subject line, message content, the five characteristics as basis for judgment, analysis, statistics and extract, enabling e-mail filtering.

This article describes a method of junk email filtering based on message content.

3.1. Junk Email Filtering Based on Message Content
Junk email filtering is to divide emails into two types, junk mails and legitimate mails, filtering out the junk mails, "spam".

E-mail classification is critical to spam filtering in order to ensure effective filtration, content-based spam filtering focuses on the content contained in e-mail as a research object, general classification algorithm begin with some known spam training as samples, extracting spam features, and then constructing a spam filterer, by which new messages are analyzed, judged, and legitimate mail is distinguished from spam and spam filtering is achieved.

Content-based spam filtering typically include mail collection, mail management and mail filtering and other steps. Generally speaking the e-mail filtering consists of two stages: learning and analysis; in the learning stage, certain algorithms are used to analyze and process the collected messages (including spam and legitimate mail) to establish an appropriate classifier, and then it is applied to filter mails in the analysis phase.

Generally, the specific implementation steps are:
1. First a certain number of spam and legitimate emails are collected in order to establish two sets of spam and legitimate e-mails.
2. The appropriate classification algorithm is used for training these known spam samples, analyzing the e-mail messages, extracting features from the e-mails and collecting corresponding data.
3. A message classifier is constructed.
4. An appropriate thresholds is selected for spam judgment, by the use of established e-mail classifiers messages are classified. The flowchart of spam mail based on content filtering is showed in Figure 2.

![Figure 2. Flowchart of Spam Filter Based on Content](image)

3.2. Description of the Boosting Algorithm
Boosting algorithm [19] is an iterative algorithm, also an effective classification algorithm, by which the accuracy of distinguishing mails can be improved greatly in mail filtering, enabling legitimate mail not to be misjudged as junk mail while reducing false contracting rate.

Boosting algorithm, aiming at training sample set, learn through an iterative process, one of whose core ideas is to remain a weight distribution on training sample set. In the initial training, the weight of all samples is equal, \(1/n\), after each iteration, the weight of failed samples is increased, and effect of the weak learning machine is strengthened for those difficult training
samples. Training and learning process, the process of structure and classification of classifier together determine the accuracy of spam filtering, therefore the various parameters used in the process should be selected reasonably, which in the actual operation and testing, should be adjusted timely, and ultimately a relatively reasonable value is determined, so as to reduce false positives to a minimum as much as possible. It is proved that as long as the error rate of each prediction function is less than 0.5 (i.e. better than random guessing), the prediction accuracy of the final prediction functions are often high, so in actual application, according to the actual situation a better classifier can be obtained by means of test and others, so the accuracy to distinguish legitimate messages is improved, reducing spam “slipping away.”

Boosting algorithm is described as follows:

1. Initial weight distribution of the training samples
   For each \( U(v, u_i) \), where \( u_i \in U \), \( v_i \in V = \{ +1, -1 \} \), \( i = 1, 2, \ldots, n \), in addition, weight-distributing of training samples in iteration is \( D_t(i) \), the predictive function generated in \( t \) iteration is represented by \( h_t \).

2. For \( t = 1 \) to \( N \) // iterate the following procedure, where \( N \) is the number of iterations, typically an experience value.
   1) Boosting algorithm is called, where \( D_t \) is the parameter Boosting algorithm;
   // \( D_t \) is weight of \( t \) round of cycle.
   2) prediction function is obtained: \( h_t : U \rightarrow V = \{ +1, -1 \} \);
   3) the error rate \( e_t \) is calculated:
   \[
   e_t = \sum _{(u,v) \in U} D_t(i) \\
   \]
   4) \( \alpha_t = \frac{1}{2} \ln \left( \frac{1 - e_t}{e_t} \right) \in R \\
   // \alpha_t \) is the weight value of \( h_t \) in \( t \) round.
   5) according to the error rate, the actual weight value of the training samples is updated:
   \[
   D_{t+1}(i) = \frac{D_t(i) \exp(\alpha_t h_t(u_i))}{Z_t} \\
   Z_t = \sum _{(u,v) \in U} D_{t+1}(u, v) = 1 \\
   \]
   6) at the end of the cycle, the classifier is obtained ultimately
   \[
   H(u) = \text{sign} \left( \sum _{t=1} D_t h_t(u) \right) \\
   \]

4. Evaluation of Spam Filtering Based on Boosting Algorithm
   In the process of spam filtering, the filterer divides emails into two categories: spam and legitimate messages, so it may cause two corresponding classification error in spam filtering process: mistaking the legitimate messages as spam, or taking spam as legitimate mail. In the former case, users’ important mail is lost, leading to serious consequences; in the latter case, a lot of trouble is caused for users. Therefore, the consequences of misjudgment caused by these two vary greatly, the cost of misjudging a legitimate e-mail is much greater than slipping a junk
mail away, which is an important reason why many users do not want to use spam filtering devices. Thus, in the spam filtering process, legitimate messages should not to be misjudged as spam as much as possible, but at the same time spam false positives should be reduced as much as possible to improve the accuracy of classification.

In this paper, indexes in the field of information retrieval and text categorization are used [20], constructing a spam filter evaluation system to evaluate the effect of spam filters. For the convenience of description, variables are defined as follows: suppose there are \( S \) messages in test set, \( C_{hs} \) is the number of legitimate messages mistaken as spam, \( N_h \) is the total number of legitimate messages, \( C_{sh} \) is the number of spam misclassified as legal mail, \( N_s \) is the total number of spam messages, then \( S = N_h + N_s \).

Evaluation is defined as follows:

1. Error rate \( E \)

\[
E = \frac{C_{hs} + C_{sh}}{N_h + N_s} = \frac{C_{hs} + C_{sh}}{S}
\]

(2)

Error rate \( E \) reflects how the e-mails are misjudged, including two false positives rate: spam misclassified as legitimate messages, legitimate messages misclassified as spam.

2. False negative rate \( M \)

\[
M = \frac{C_{sh}}{N_s}
\]

(3)

False negative rate \( M \) refers to the ratio of spam misclassified as legitimate messages, which reflects the spam recognition of the filterer.

3. False rate \( F \)

\[
F = \frac{C_{hs}}{N_h}
\]

(4)

False rate \( F \) reflects the legitimate messages incorrectly identified as spam, in spam filtering process, which should be avoided from happening.

4. Recall rate \( R \)

Recall rate \( R \) includes spam recall \( R_s \) and legitimate messages recall \( R_h \).

\[
\begin{align*}
R_s &= 1 - M = \frac{N_s - C_{sh}}{N_s} \\
R_h &= 1 - F = \frac{N_h - C_{hs}}{N_h}
\end{align*}
\]

(5)

Recall rate \( R \) reflects how spam is filtered out and legitimate messages get through. Spam recall \( R_s \) is spam detection rate, which reflects the spam filterer's ability to find spam, spam recall rate \( R_s \) higher, the less spam "slipping away"; legitimate mail recall \( R_h \) is the rate of legitimate mail getting through, which reflects the spam filterer's ability to let legitimate mail through, the recall rate \( R_h \) higher, the less legitimate messages misjudged.

5. Precision rate \( P \)
The precision rate $P$ includes spam precision rate $P_s$ and legitimate e-mail precision rate $P_h$.

\[
\begin{align*}
P_s &= \frac{N_s - C_{sh}}{N_s - C_{sh} + C_{hs}} \\
P_h &= \frac{N_h - C_{hs}}{N_h - C_{hs} + C_{sh}}
\end{align*}
\]  

(6)

Precision rate $P$ reflects how the emails are sorted out, the precision rate $P_s$ is the rate of spam correctly sorted out, which reflects the filterer’s ability to find spam, the higher spam precision rate $P_s$ means fewer legitimate messages misjudged as spam; legitimate e-mail precision rate $P_h$ is the rate of legitimate messages detected correctly, which reflects the filterer’s ability to find legitimate e-mail, the higher legitimate e-mail precision rate $P_h$ means fewer spam mistaken as legitimate messages.

6. Accuracy rate $A$

\[
A = \frac{(N_s - C_{sh}) + (N_h - C_{hs})}{S} = 1 - \frac{C_{sh} + C_{hs}}{S} = 1 - E
\]

(7)

Accuracy rate $A$ is the for all mail (including spam and legitimate mail) judgments, which reflects the filterer’s ability to determine spam as spam, and legitimate messages as legitimate messages.

7. $F$ Value

$F$ Value includes spam $F$ Value $F_s$ and legitimate messages $F$ Value $F_h$.

\[
\begin{align*}
F_s &= \frac{2P_s \cdot R_s}{P_s + R_s} \\
F_h &= \frac{2P_h \cdot R_h}{P_h + R_h}
\end{align*}
\]  

(8)

$F$ Value is the harmonic mean of precision rate $P$ and recall rate $R$, integrating the correct rate $P$ and recall $R$ into one evaluation. The precision rate $P$ and recall rate $R$, from different angles, reflect the filterer’s classification quality; in general, these two indexes are complementary, the accuracy rate $P$ being improve will lead to lower recall rate $R$, and vice versa. Spam $F$ Value $F_s$ the harmonic mean of spam accuracy rate $P_s$ and spam recall rate $R_s$, legitimate messages $F$ Value $F_h$ is the harmonic mean of legitimate e-mail precision $P_h$ and legitimate messages recall rate $R_h$.

8. Total Cost Ratio ($TCR$)

In spam filtering, it is not desirable that legitimate messages are incorrectly identified as spam. In order to express spam filterer’s cost in different situations, Androutsopoulos I and others proposed the concept of Total Cost Ratio, $TCR$ [21], $TCR$ is defined as:
\[
TCR = \frac{N_s}{\lambda C_{hs} + C_{sh}}
\]

\(\lambda\) is the ratio of legitimate messages incorrectly identified as spam and spam incorrectly identified as spam, i.e. \(\lambda = C_{hs} / C_{sh}\).

\(TCR\) reflects the filterer’s performance; the higher \(TCR\) is, the lower the loss of the filterer is, and the better the filterer’s performance is.

All of the above performance indexes, from different angles, evaluate a spam filterer, and therefore an evaluation system can be constituted by the above indexes, according to the experimental test data, the filterer can be evaluated comprehensively, classification effect of the filterer based on Boosting algorithm are judged.

5. Experiment Results and Analysis

For the experimental environment, the hardware used in this paper is Sun servers; software is the Solaris Operating System to build a dedicated mail server. Multiple experiments are conducted with the spam filterers based on Boosting algorithm and the spam filterers based on the traditional Bayesian algorithm (in each experiment 800 different e-mails are collected, among which is 300 legitimate e-mails, spams 500), the experimental data strike a mean to obtain results closer to the real situation. A number of simulation experiments of information filtering are conducted with algorithm based on Boosting algorithm and the traditional Bayesian algorithm [22], the data are shown in Table 1.

Table 1. Experimental Results of Filtering Spam using Boosting Algorithm and Traditional Bayesian Algorithm

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>(C_{hs})</th>
<th>(N_h)</th>
<th>(C_{sh})</th>
<th>(N_s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosting</td>
<td>6.82</td>
<td>300</td>
<td>16.56</td>
<td>500</td>
</tr>
<tr>
<td>Bayesian</td>
<td>16.38</td>
<td>300</td>
<td>32.78</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 1 is classification of test data obtained by the junk mail filterer based on Boosting algorithm on 800 messages, including 300 legitimate messages and 500 spam messages, the filtration results are 310.28 legitimate mails and 489.72 junk mails; while filtering results based on the traditional Bayesian algorithm is 316.40 legitimate messages and 483.60 junk mails.

We have already established an evaluation system composed of \(E\), \(M\), \(F\), \(R\), \(P\), \(A\), \(F Value\), \(TCR\), the eight performance indicators; by means of data in Table 1 and the performance indicators of the evaluation system, we obtained performance indicators statistical data comparison between the two filterer based on Boosting algorithm and traditional Bayesian algorithm as shown by Table 2.

Table 2. Comparison between the Experimental Results Based on the Two Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>(E)%</th>
<th>(M)%</th>
<th>(F)%</th>
<th>(R_d)%</th>
<th>(R_h)%</th>
<th>(P_d)%</th>
<th>(P_h)%</th>
<th>(A)%</th>
<th>(F Value)</th>
<th>(F_{Rs})%</th>
<th>(F_{Rh})%</th>
<th>(TCR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosting</td>
<td>2.92</td>
<td>3.31</td>
<td>2.27</td>
<td>96.69</td>
<td>97.73</td>
<td>98.61</td>
<td>94.65</td>
<td>97.08</td>
<td>97.64</td>
<td>96.17</td>
<td>25.81</td>
<td></td>
</tr>
<tr>
<td>Bayesian</td>
<td>6.15</td>
<td>6.56</td>
<td>5.46</td>
<td>93.44</td>
<td>94.54</td>
<td>96.61</td>
<td>89.64</td>
<td>93.86</td>
<td>95.00</td>
<td>92.02</td>
<td>12.21</td>
<td></td>
</tr>
</tbody>
</table>

From the data in Table 2 it can be clearly drawn that Boosting algorithm based spam filter evaluation on \(E\), \(M\) and \(F\) were significantly lower than those of traditional Bayesian algorithm-based spam filtering device, while Boosting algorithm-based spam filterer’s \(R\), \(P\), \(A\) and \(F Value\) were significantly higher than the four evaluation by the spam filters based on the traditional Bayesian algorithm, especially as for the \(TCR\), Boosting algorithm-based spam filterer is much higher than the traditional Bayesian algorithm-based spam filters.
6. Conclusion

The above comparison of experimental data shows that junk mail filterer based on Boosting algorithm is significantly better than the junk mail filterer based on traditional Bayesian algorithm, the former is able to achieve an excellent spam filtering function, so the effectiveness of Boosting algorithm in junk mail filtering is proved. Furthermore, a more suitable kernel function and its parameters will be selected in our future research so as to improve the identification rate of the filterer.

Acknowledgements

This work is supported by the Natural Science Foundation of Heilongjiang Province, China under grant No. F201331.

References