This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30029158

©2009 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright: 2009, IEEE
MVGL Analyser for Multi-classifier Based Spam Filtering System

Md Rafiqul Islam
School of Information Technology, Deakin University
Melbourne, Australia.
rislam@deakin.edu.au

Wanlei Zhou
School of Information Technology, Deakin University
Melbourne, Australia.
wanlei@deakin.edu.au

Morshed U Chowdhury
School of Information Technology, Deakin University
Melbourne, Australia.
muc@deakin.edu.au

Abstract - In the last decade, the rapid growth of the Internet and email, there has been a dramatic growth in spam. Spam is commonly defined as unsolicited email messages and protecting email from the infiltration of spam is an important research issue. Classifications algorithms have been successfully used to filter spam, but with a certain amount of false positive trade-offs, which is unacceptable to users sometimes. This paper presents an approach to overcome the burden of GL (grey list) analyzer as further refinements to our multi-classifier based classification system (Islam, M. and W. Zhou 2007). In this approach, we introduce a “majority voting grey list (MVGL)” analyzing technique which will analyze the generated GL emails by using the majority voting (MV) algorithm. We have presented two different variations of the MV system, one is simple MV (SMV) and other is the Ranked MV (RMV). Our empirical evidence proofs the improvements of this approach compared to the existing GL analyzer of multi-classifier based spam filtering process.

Key Words- Spam, GL analyzer, FP, MVGL, Multi-classifier.

I. INTRODUCTION

The Internet has rapidly become an integral part of everyday life and its usage is expected to continue growing (Miniwatts 2007). The usage of Internet email has emerged as one of the primary tools of communication, intended for idea and information exchange throughout the world (Stanford 2007). However, its rapid adoption has also left it susceptible for misuse and abuse. Along with the growth of the Internet and email, there has been a dramatic growth in spam in recent years (Barracuda 2007).

The increasing volume of spam is rapidly becoming a serious problem (Research 2003; Claburn 2005; Symantec 2006; Barracuda 2007). Effectively dealing with unwanted email is important not only for cost savings but also to protect Internet users. Spam filtering is able to control the problem in a variety of ways. Identification of spam and its removal method from the email delivery system allows endusers to regain a useful means of communication. However, the key challenge is that spam is difficult to define, as what is spam to one person is not necessarily spam to another. These result is not only spam passing through the filter, but also in raising false positives alarms where legitimate emails may be blocked. If there are no effective anti-spam mechanisms, spam will inundate network systems, cause significant loss of productivity, steal bandwidth (Symantec 2006), block legitimate email correspondence, and still be there tomorrow.

This paper presents an approach to minimize the analysing complexity of GL analysing technique of our previous multi-classifier based spam filtering model (Islam, M. and W. Zhou 2007). Using this approach, the system not only reduces the burden of the analyser but also overcome the limitations of human interaction of the existing GL analyser. The experimental results show the significant performance of this approach compared to the existing system, in particular minimizing the analysing complexity. The rest of the paper organizes as follows: Section 2 describes the minimizing technique using MVGL along with different variation. Section 3 presents the empirical evidence of the proposed approach and finally section 4 conclude the paper with the direction of future work.

II. MVGL TECHNIQUE

In the spam filtering model based on multi-classifier classification technique, presented in (Islam, M. and W. Zhou 2007 a & b), the grey list (GL) analyzer of multi-classifier model introduces some added complexity in terms of processing time and memory overhead. It also depends to some extent on human interaction to get the final decision about the GL emails (Islam, M. and W. Zhou 2007a). To overcome this shortcoming we have introduced the MVGL technique. The concept of the MV method is the simplest of all combinatorial functions(Lam and Suen 1997; De Stefano, della Cioppa et al. 2002; Bhattacharya and Chaudhuri 2003; Hong, Chengde et al. 2007). It selects the relevant class prediction by polling all the classifiers to see which class is the most acceptable. Whichever class gets the highest support is selected. This method is particularly successful when the classifiers involved output integer selections. The following section outlines the details of MVGL technique.

A. MVGL for N-Classifier

The MV technique has attracted much attention in the case of the multi-classifier classification system for its simplicity and high level of accuracy and robustness which can be archived in appropriate incidents (Hong, Chengde et
The impressive performance of MV has been demonstrated in various applications in biometric recognition such as handwriting recognition (Bhattacharya and Chaudhuri 2003) and pattern recognition (Lam and Suen 1997). The MV system has quite a number of variations in terms of application and methodology (Kuncheva 2002; Rahman and Fairhurst 2003), although the underlying principle is the same. The viewpoint of the variations of MV is based on two basic strategies:

- The pronouncement will be accepted if the majority of the classifiers have the same opinion, without considering the trustworthiness/ranking of the classifiers.
- The pronouncement will be accepted if the most competent classifiers have the same opinion, without giving the consensus of the majority classifiers.

Both strategies are useful and can achieve fruitful performance but need to require careful integration within the decision making process. In our proposed approach we introduce both of the strategies to minimize the complexity of the analyzer.

There are different distinctions of using the MV techniques presented in (Rahman, Alam et al. 2002) such as weighted MV, restricted MV, Enhanced MV and ranked MV. But the performance of the various approaches of the MV techniques is directly related to their design emphasis. However, emphasis has been given here to assess how a consensus can be reached given the often conflicting opinions of classifiers in the multi-classifier classification environment. In our proposed approach, we have considered couple of variations; one is simple MVGL (SMVGL) and other is ranked MVGL (RMVGL) techniques. The following section outlines the details of SMVGL and RMVGL technique.

- **Simple MVGL (SMVGL)**

  In the multi-classifier classification environment presented in (Islam and Zhou, 2007a), an adaptive section will collect all the output of the classifiers and categories them in three different classes based on their consensus. The unique predictions from all the classifiers will be sent to the corresponding TP (true positive) and TN (true negative) mailboxes and the mixed-predictions will be sent to the GL mailbox, as shown in figure 1. Analyzing the GL outputs of the classifier, the SMVGL technique will be used when the decision is assigned to a class label for which there is a consensus, or when the majority (that is more than half) of the classifiers agrees on a unique class label. Otherwise the GL outputs will go to the RMVGL technique. The following figure 1 shows the SMVGL technique for the N-classifier, of our multi-classifier based spam filtering model.

![Figure 1: SMVGL for N-classifier.](image-url)
Let \( \text{Pred}(i, n) \) denote the \( N \) number of classifiers prediction of \( i \). Each classifier's prediction is represented by binary expressions (1 or -1) with 1 indicating the classifier is correct predicted and -1 indicating it is incorrect or misclassified. If there are \( N \) classifiers \( C_1, C_2, \ldots, C_N \), \( C_1 \) is the \( LS \) (least significant) and \( C_N \) is the \( MS \) (most significant) then there are \( 2^N - 2 \) possible combinations of diverse classification that is correct/incorrect classifications which are treated as GL product for \( N \) classifiers as described before.

From figure 1, it has been shown that the adaptive section categorized the classifiers output into three different categories TP, TN and GL. The category GL comes to SMV technique for counting individual prediction to support a final decision. After getting the final decision the emails will be sent to the corresponding mailbox that is into the TP or TN. Mathematically we can represent the MV algorithm using the max function for legitimate outputs (Islam, M. and W. Zhou 2007a & 2007b).

\[
\int (C_1, C_2, \ldots, C_n) \Rightarrow \prod_{i=1}^{n} C_i^{l} + \prod_{i=1}^{n} C_i^{s} + \max_i \left( \sum_{x=1}^{k} \prod_{j=1}^{p} C_j^{s} \prod_{k=1}^{q} C_k^{l} \right)
\]

Where \( C_1, C_2, \ldots \) are the classifiers, \( C^l \) and \( C^s \) are the legitimate and spam output of the \( i^{th} \) classifiers respectively. The total numbers of GL emails are (from figure 1):

\[
\text{GL} \Rightarrow \int (C_1, C_2, \ldots, C_n) \Rightarrow \max_i \left( \sum_{x=1}^{k} \prod_{j=1}^{p} C_j^{s} \prod_{k=1}^{q} C_k^{l} \right)
\]

\[
\Rightarrow \max_i \left( C_1^{l}C_2^{l} \ldots C_{m-1}^{l}sC_{m}^{s} + C_1^{s}C_2^{l} \ldots C_{m-1}^{l}sC_{m}^{l} + \ldots + C_1^{s}C_2^{s} \ldots C_{m-1}^{s}sC_{m}^{l} \right)
\]

\[
(1)
\]

Let \( b(i; k) \) denote the \( K \) bit binary representation of integer \( i \). For determining \( K \), there are two possible cases –

Case 1: Odd number of classifiers (\( N \)) the value of \( K \) is \((N+1)/2\)

Case 2: Even number of classifiers (\( N \)) the value of \( K \) is \((N/2)+1\)

If the \( K \) represents the positive label for a particular email then the analyzer will be treated as legitimate and vice versa for spam. According to the figure 1, the MV algorithm will count the individual classifier predictions for every email and the pronouncement will be accepted for legitimate emails if majority function \( f_{\text{max}}(\text{Pred}(i,n)) \) is positive, that is, for legitimate emails the majority function is-

For case 1:

\[
f_{\text{max}}(i, l) = (l \mid \text{bit}(i,n) > n/2) \text{ and }
\]

For case 2:

\[
f_{\text{max}}(i, l) = (l \mid \text{bit}(i,n) > n/2+1)
\]

where \( n \) is the number of classifiers

\[
(3)
\]

\[
(4)
\]

- **Ranked MVGL**

It is an enhancement of SMVGL. In this approach the decision of the individual classifier along with the rank factor of the classifier will be multiplied to get the final decision. The rank factor is the comparative competence of the collaborating classifiers. The higher the competences value the higher of its rank factor. Let \( R_iC_i \) be the rank factor of the \( i^{th} \) classifier expressing the comparative \( k \) class competence of the classifiers be denoted by \( d_{ik} = R_iC_i \) with \( i=1..N \) and \( k=0..1 \). So the final decision \( D_{\text{final}} = \max \text{ of } d_{ik} \).

For an even number of classifiers, one of the probabilities could arise in the case of tie that is \( k=n/2 \). In that case the system will consider the classifier ranking. The system will then count the rank of the classifier simultaneously and, based on the rank majority voting (RMV), the email will be determined accordingly. The figure 2 illustrates the RMV system for GL emails.

In this approach the GL analyzer will determine the rank of the classifier based on the precision value along with classification predication. In the case of a tie, the analyzer will evaluate the rank value of each classifier. If the rank value is higher for the legitimate case, then the analyzer will predict as positive otherwise negative and will send the email to the corresponding folder. This process will be initiated only in the case where the \( k \) value is \( n/2 \).
Let $C_1, C_2, C_3, ..., C_N$ are the classifier and $R_1, R_2, R_3, ..., R_N$ are the corresponding rank values. For an even number of classifiers when the probability factor of the collective consensus is 0.5, then the system will determine the legitimate email that is-

$$D_{\text{final}} = \max \, \text{of} \, R_i C_{il}.$$

(5)

And the rank factor of the classifier will be determined based on the precision of the classifier, that is -

$$P_i = \frac{\text{true positive }_i}{(\text{true positive }_i + \text{false positive }_i)}$$

(6)

If the precision is higher, then the rank of the classifier is higher and vice versa. The reason behind this philosophy is that our objective is to reduce the false positive, so higher precision means lower false positive.

---

**III. EXPERIMENTAL RESULTS**

In our experiment, we have used the public data sets PUA1-2-3 (Androutsopoulos, Koutsias et al. 2000) and converted the data sets based on our experimental design and environment. Firstly we have encoded the whole data sets, both train and test sets, then indexed every email for test data sets and finally recorded the output according to the index value.

A. **Experimental setup**

Programming Language: Matlab-7.1

Basic Steps:

- Encode the email content.
- Collect all the individual emails and make email data sets (Matlab format): one for training data sets and another for test data sets
- Index the test data sets
- Train the classifiers using training data sets
- Classify the test data based on index value

The figure 3 shows the comparison of proposed MVGL approach with our previous GL analyser (Islam and Zhou, 2007). It is to be noted that, we have only used the MV system in only for GL emails of the system, not for the others. It has been shown that the MVGL technique is not
outperforms compare to GL analyser performance in terms of accuracy but it reduces the complexity except the data set 3. However, on an average the performance is convincing.

Figure 3: The performance comparison of MVGL and GL analyser.

Figure 4: The performance comparison of MVGL and Multi-classifier classification

B. ROC Report

Table 1 shows the comparative ROC report of MVGL and multi-classifier technique. Four ROC estimations are used to compare the performance such as AUC, AUC standard error (StdErr), 95% confidence interval and P-value. It has been shown that the AUC estimation for multi-classifier is better than that of MVGL technique; however the AUC StdErr and CI value is better in MVGL technique. Figure 4 shows the ROC curve for both the techniques which proofs the same indication as shown in Table 1.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>ROC Estimation</th>
<th>Multi-Classifier</th>
<th>MVGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUD (1-6)</td>
<td>AUC 0.97222 0.96506</td>
<td>0.97222 0.96506</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUC (StdErr) 0.01936 0.01783</td>
<td>0.01936 0.01783</td>
<td></td>
</tr>
<tr>
<td></td>
<td>95% of CI 0.81421 0.85803</td>
<td>0.81421 0.85803</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P-Value (&lt;0.001) (&lt;0.001)</td>
<td>(&lt;0.001) (&lt;0.001)</td>
<td></td>
</tr>
</tbody>
</table>

IV. CONCLUSION & FUTURE WORK

This paper presents the techniques for reducing analysing techniques of our previous multi classifier classification technique (Islam and Zhou, 2007). We have investigated the MVGL approach to reduce the analysing complexity of GL analyser and found better performance. It has been shown that MVGL technique reduces the analysing complexity of GL analyser. However, the MVGL technique sometimes reduces the accuracy compared to the GL analyser. We are investigating and working on it and will explore it in our future work.

REFERENCES:


De Stefano, C., A. Della Cioppa, et al. (2002). An adaptive weighted majority vote rule for combining multiple


