Feature Subset Selection in E-mail Spam Detection

Amir Rajabi Behjat, Universiti Technology MARA, Malaysia

“IT Security for the Next Generation”
Asia Pacific & MEA Cup, Hong Kong
14-16 March, 2012
Feature Subset Selection in E-mail Spam Detection

Introduction
Problem Statement
Objective
Related Works
Proposed Architecture
Result and Discussion
Conclusion
Introduction

Today, e-mails have become a very common and convenient medium for our daily communications. Spam is an unsolicited commercial or bulk e-mails, or an uninterested e-mails. It has created a significant security problem for end users and big organizations through carrying malicious virus and phishing scams.

There are many solutions available to act as a model for automatic spam detection. Most techniques are based on Machine Learning and data mining techniques such as classification, clustering and Genetic Algorithm that classify spam emails and select their relevant features (Sang, 2010).
Problem Statement

- Lack of useful and relevant features that can distinguish between spam and non-spam emails efficiently (Bilal, 2005).
- Increase data dimensionality that decreases accuracy (Ren, 2006).
- The reduction of classification accuracy within the current methods (Burim, 2005).
- A high growth of false positive in different spam detection systems (Ching, 2010).
Objective

- To find related and relevant features to distinguish between spam and non-spam emails and decrease data dimensionality.

- To increase classification accuracy that shows the accuracy of detection system and MLP classifier.

- To decrease false positive using GA as a feature selector and MLP classifier.
## Related Works

<table>
<thead>
<tr>
<th>Author and Year</th>
<th>Proposed System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruan and Tan (2007)</td>
<td>Classified spam and legitimate emails based on support Vector Machine (SVM) and select more relevant features using Genetic Algorithm.</td>
</tr>
<tr>
<td>Matthew and Chung Keung (2009)</td>
<td>Selected features as a mathematical method and classified them based on k-NN and Naïve Bayes. The results of proposed detection system show an accuracy near to 98%.</td>
</tr>
<tr>
<td>Gaurav and Shashikala (2010)</td>
<td>Detected spam emails based on different techniques such as analyzing email contents by spam keywords knowledge base, finding sender of emails, and cross validation without using feature selection. The lack of feature selection and high computational costs decrease accuracy of classifier.</td>
</tr>
<tr>
<td>Lee, Kim and Park (2010)</td>
<td>Developed an optimal spam detection model based on Random Forest (RF). This model is able to optimize RF parameters and select relevant features simultaneously.</td>
</tr>
<tr>
<td>Ruan and Tan (2010)</td>
<td>Selected features of spam emails such as words and symbols by artificial Immune system to detect spam emails based Artificial Neural Network classifier. The result of this study showed 99% accuracy and 0.2% error rate.</td>
</tr>
</tbody>
</table>
Detection System Architecture
Genetic Algorithm

- GA finds an optimal binary vector, where each bit is associated with a feature.
  - If the $ith$ bit of this vector = 1, then the $ith$ feature is allowed to participate in classification;
  - if the bit is = 0, then the corresponding feature does not participate.

- GA method applied in this study decreases the number of irrelevant features and high dimensionality. Thus, the number of features are decreased from 156 to 78.
The Fitness function of GA
Multi-Layer perceptron (MLP)

- Multi-layer perceptron is arranged in different layers, namely input, hidden and output layers. In fact, these layers are organized to minimize appropriate error functions and increase the accuracy of classification.
- This classifier not only increase the accuracy, but increase the speed of detection process.
Dataset

Two benchmark corpora used to test our proposed email spam detection system based MLP classifier are the SpamAssassin corpus and LingSpam corpus.

<table>
<thead>
<tr>
<th>Name of Database</th>
<th>Spam</th>
<th>Non-spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>LingSpam</td>
<td>481</td>
<td>2412</td>
</tr>
<tr>
<td>SpamAssassin</td>
<td>3797</td>
<td>6954</td>
</tr>
</tbody>
</table>
### Performance Measurement

<table>
<thead>
<tr>
<th>Measure</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>( \frac{TP+TN}{TP+FP+FN+TN} )</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>( \frac{TP}{TP+FP} )</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>( \frac{TP}{TP+FN} )</td>
</tr>
</tbody>
</table>

TP = spam emails and identified as spam  
FN = spam emails but identified as non-spam  
TN = non-spam emails and identified as non-spam  
FP = non-spam emails but identified as spam
Table 1 illustrates the performance of GA based feature subset selection with MLP classifier in the spam detection system.

10 runs were used for each dataset to estimate the classifications.

Other techniques from previous studies are compared to the proposed technique i.e MLP

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy (%)</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Discriminant</td>
<td>98.55</td>
<td>91.49</td>
<td>99.77</td>
</tr>
<tr>
<td>SVM</td>
<td>97.09</td>
<td>97.38</td>
<td>97.02</td>
</tr>
<tr>
<td>BP Neural Network</td>
<td>99.13</td>
<td>95.84</td>
<td>98.93</td>
</tr>
<tr>
<td>NB</td>
<td>96.41</td>
<td>81.10</td>
<td>96.85</td>
</tr>
<tr>
<td>SVM-IG</td>
<td>96.85</td>
<td>81.90</td>
<td>99.0</td>
</tr>
<tr>
<td>MLP - LingSpam</td>
<td>99.83</td>
<td>93.75</td>
<td>100</td>
</tr>
<tr>
<td>MLP - SpamAssassin</td>
<td>99.81</td>
<td>100</td>
<td>99.48</td>
</tr>
</tbody>
</table>
Cont. Result

Performance of Spam Detection System

- Linear Discriminant
- SVM
- BP Neural Network
- NB
- Linger-V
- SVM-IG
- MLP
The Genetic Algorithm (GA) as a feature selection algorithm is discussed in this study. This algorithm like an automated heuristic method selects significant features of spam emails to decline data-dimensionality and increase MLP classifier.

The results of this paper show the performance near to 100% and 0.01 false positive rate. Integrated experiments of this paper are based on two benchmark corpora LingSpam and SpamAssassin to indicate the strength of GA in feature selection and the ability of MLP classifier for generating high accuracy and decrease the time of computing.
Thank You

Amir Rajabi Behjat, Universiti Technology Mara, Malaysia

“IT Security for the Next Generation”
Asia Pacific & MEA Cup, Hong Kong
14-16 March, 2012