A Novel Approach For Generating Rules For SMS Spam Filtering Using Rough Sets

Ashima Wadhawan, Neerja Negi

Abstract: Spam is defined as unwanted commercial messages to many recipients. Email Spamming is a universal problem with which everyone is familiar. This problem has reached to the mobile networks also now days to a great extent which is referred to as SMS Spamming. A number of approaches are used for SMS filtering like blacklist-white list filter, Content based filter, Bayesian filtering, checksum filter, heuristic filter. The most common filtering technique is content based spam filtering which uses actual text of messages to determine whether it is spam or not. Bayesian method represents the changing nature of message using probability theory. Bayesian classifier can be trained very efficiently in supervised learning. We have used a new mathematical approach Rough set Theory. Rough Set Theory is a new methodology which is used to cluster the objects of a decision system with a large data set. In this dissertation, the Naïve Bayes and the RST method are implemented.

Index Terms: Bayesian Filtering, Classification, Checksum Filter, Content Based Filtering, Heuristic Filtering, Rough set, SMS Spam Filtering

1 INTRODUCTION

With the development of Internet and the rapid increase of network bandwidth, spam mail or also call Unsolicited Commercial Email(UCE) is increasingly becoming a great problem today. One of the reasons for the exponential growth of spam is where the email which has provides a cheap and neat instantaneous mode of communication world-wide. Spam has caused some serious problem that alert email user nowadays Spam can be defined as unsolicited (unwanted, junk) electronic message in which the number of recipient has not granted the permission for it to be sent[1].Spam is distributed in a widely variety of forms including email spam, instant messaging spam, SMS spam, image spam. SPAM stands for Short pointless annoying message that describe sort of things. SMS has certain characters that are different from mails. A mail consists of certain structured information such as subject, mail header, salutation, sender's address etc. but SMS lacks such structured information. These make the SMS classification task much difficult. This situation makes the necessity for developing an efficient SMS filtering method.

2 RELATED WORK

Before 1990, some Spam prevention tools began to emerge in response to the Spammers who started to automate the process of sending Spam email. The first Spam prevention tool has used simple approach, based on language analysis by simply scanning emails for some suspicious senders or phrases like “click here to buy" and "free of charge”. In late 1990s, blacklisting and white listing methods were implemented at the Internet Service Provider (ISP) level. However, these methods suffered from some maintenance problems. There are many efforts underway to stop the increase of Spam that plagues almost every user on the mobile network. Various techniques have been used to filter the Spam messages. Bayesian [1] classifier is a simple probabilistic classifier. Its main advantage is that naive Bayes classifiers can be trained very efficiently in a supervised learning. Bayesian classifiers are used for parameter estimation in numerous practical applications. In supervised learning, the parameters are estimated by Maximum Likelihood Estimation (MLE) method. Decision Tree [2] is one of the most famous tools of decision-making theory. Decision tree is a classifier in the form of a tree structure that shows the reasoning process. Rough Sets[3] is a new methodology which is used to cluster the objects of a decision system with a large data set. An Information System is represented as IS = (U, A), where, U is the Universal set of objects and C is a set of condition attributes. Here, we deal with a Decision System, which is represented as DS = (U, AU(d)), where d is a decision attribute. An Indiscernibility Relation is defined on a subset B of A as RB = {(x, y)εU x U | a(x) = a(y), for all aεA}, where a(x) is the value of object x for attribute a. The set U is partitioned into different sets based on the decision classes of a decision attribute and the equivalence classes are obtained based on B. Let there be k decision classes, d1, d2, …., dk. The equivalence classes based on the decision attribute, d, are represented as [U]dk. Clearly, [U]d is a subset of U. Let [U]d be denoted as X i.e. X ⊆ U. Let the equivalence classes obtained from the Indiscernibility relation be denoted by [X]B. There is no work done for Rough set SMS Spam filtering yet and it is much more necessary to start the work.

3 PROPOSED WORK

The proposed system framework contains four steps: Data set, preprocessing, Bayesian filtering classification, Decision rules.

3.1 DATA SET AND PREPROCESSING:

Firstly take a data set. The purpose of preprocessing is to transform messages in SMS to in a uniform format. It can take some attributes and also taken a corpus set(training set) if every attribute which you have taken that can match with every message of the corpus set then that consider 1 otherwise 0.
### 3.2 BAYESIAN FILTERING CLASSIFICATION

<table>
<thead>
<tr>
<th>sms</th>
<th>Se_know</th>
<th>we_bms</th>
<th>Longst</th>
<th>thanks</th>
<th>Congr</th>
<th>win</th>
<th>free</th>
<th>sorry</th>
<th>urgent</th>
<th>priv</th>
<th>please</th>
<th>finall</th>
<th>serv</th>
<th>offe</th>
<th>great</th>
<th>oops</th>
<th>rem</th>
<th>nder</th>
<th>cal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**MEAN**

<table>
<thead>
<tr>
<th>spam</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.75</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ham</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
We wish to determine which posterior is greater ham or spam for classification as spam the posterior is given by

\[
posterior_{\text{spam}} = P(\text{spam})p(\text{sender\_known}|\text{spam})p(\text{webmsg}|\text{spam})p(\text{longstring}|\text{spam})p(\text{thanks}|\text{spam})p(\text{congratulations}|\text{spam})p(\text{win}|\text{spam})p(\text{free}|\text{spam})p(\text{sorry}|\text{spam})p(\text{urgent}|\text{spam})p(\text{private}|\text{spam})p(\text{please}|\text{spam})p(\text{finally}|\text{spam})p(\text{service}|\text{spam})p(\text{offer}|\text{spam})p(\text{great}|\text{spam})p(\text{oops}|\text{spam})p(\text{reminder}|\text{spam})p(\text{call}|\text{spam})
\]

\[
posterior_{\text{ham}} = P(\text{ham})p(\text{sender\_known}|\text{ham})p(\text{webmsg}|\text{ham})p(\text{longstring}|\text{ham})p(\text{thanks}|\text{ham})p(\text{congratulations}|\text{ham})p(\text{win}|\text{ham})p(\text{free}|\text{ham})p(\text{sorry}|\text{ham})p(\text{urgent}|\text{ham})p(\text{private}|\text{ham})p(\text{please}|\text{ham})p(\text{finally}|\text{ham})p(\text{service}|\text{ham})p(\text{offer}|\text{ham})p(\text{great}|\text{ham})p(\text{oops}|\text{ham})p(\text{reminder}|\text{ham})p(\text{call}|\text{ham})
\]

\[
p(\text{free}|\text{spam})=1/\sqrt{2\pi\sigma^2}\exp(-\frac{(x-\mu)^2}{2\sigma^2})
\]

for every attribute|spam or ham i.e we can apply the same formula...

\[
posterior_{\text{spam}}=0.5\times1*0.3678*1*1*0.3678*0.2587*1*0.3678*1*1*1*1*1*0.6870=0.0044
\]

\[
posterior_{\text{ham}}=0.5*0.3678*1*0.3678*1*1*0.3678*1*1*1*1*1*1*0.3678=0.0033
\]

so posterior_{\text{spam}} is greater than posterior_{\text{ham}}

we predict the sample is spam
3.3 DECISION RULES USING RST
RST is a mathematical tool that used to find the decision rules. It convert the data in to required format(.isf) for applying Rough set theory. It can generate the decision rules using rst

3.3.1 Approximations

<table>
<thead>
<tr>
<th>Class</th>
<th>No. of objects</th>
<th>Lower approximation</th>
<th>Upper approximation</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>674</td>
<td>665</td>
<td>684</td>
<td>0.9722</td>
</tr>
<tr>
<td>1</td>
<td>126</td>
<td>116</td>
<td>135</td>
<td>0.8593</td>
</tr>
</tbody>
</table>

3.3.2 Reduct

<table>
<thead>
<tr>
<th>#</th>
<th>Reduct</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sender Known, web msg, long String, win, free, sorry, service, offer, great, call</td>
<td>10</td>
</tr>
</tbody>
</table>

3.3.3 Core Viewer

Quality of classification

<table>
<thead>
<tr>
<th>For all condition attribute</th>
<th>0.9762</th>
</tr>
</thead>
<tbody>
<tr>
<td>For all Condition attribute in core</td>
<td>0.9762</td>
</tr>
</tbody>
</table>

Attributes in CORE

Core sender_known
Core web msg
Core longstring
Core win
Core free
Core sorry
Core service
Core offer
Core great
Core call

3.3.4 RULES
# ModLEM with Entropy
# C:\Program Files\ROSE2\examples\smsspam.isf
# objects = 800
# attributes = 19
# decision = spam
# classes = {0, 1}
# Sun May 11 14:17:38 2014
# 0
| Rule 1: | (sender_known = 1) & (sorry = 0) => (spam = 0); | [657, 657, 97.48%, 100.00%] | ![Rule 1](image1.png) |
| Rule 2: | (great = 1) => (spam = 0); | [12, 12, 1.78%, 100.00%] | ![Rule 2](image2.png) |
| Rule 3: | (sender_known = 1) & (offer = 1) => (spam = 0); | [3, 3, 0.45%, 100.00%] | ![Rule 3](image3.png) |
| Rule 4: | (sender_known = 1) & (call = 1) => (spam = 0); | [38, 38, 5.64%, 100.00%] | ![Rule 4](image4.png) |
| Rule 5: | (sender_known = 0) & (web_msg = 0) => (spam = 1); | [103, 103, 81.75%, 100.00%] | ![Rule 5](image5.png) |
| Rule 6: | (sender_known = 0) & (free = 1) => (spam = 1); | [29, 29, 23.02%, 100.00%] | ![Rule 6](image6.png) |
Rule 7: 
(sender_known = 0) & (offer = 1) => (spam = 1); [8, 8, 6.35%, 100.00%][0, 8] 
[{}, {260, 297, 368, 464, 528, 581, 637, 798}]

Rule 8: 
(service = 1) => (spam = 1); [18, 18, 14.29%, 100.00%][0, 18] 
[{}, {33, 39, 61, 94, 140, 160, 166, 189, 269, 369, 376, 416, 423, 595, 661, 739, 749, 753}]

Rule 9: 
(sender_known = 0) & (please = 1) => (spam = 1); [10, 10, 7.94%, 100.00%][0, 10] 
[{}, {25, 29, 33, 39, 52, 66, 94, 124, 160, 189}]

Rule 10: 
(sender_known = 0) & (congratulations = 1) => (spam = 1); [3, 3, 2.38%, 100.00%][0, 3] 
[{}, {251, 358, 506}]

Rule 11: 
(sender_known = 0) & (win = 1) => (spam = 1); [18, 18, 14.29%, 100.00%][0, 18] 
[{}, {12, 94, 115, 135, 168, 189, 274, 313, 320, 336, 358, 390, 506, 565, 577, 588, 718, 767}]

Rule 12: 
(long_string = 1) => (spam = 1); [4, 4, 3.17%, 100.00%][0, 4] 
[{}, {16, 20, 48, 711}]

Rule 13: 
(sender_known = 0) & (web_msg = 1) & (long_string = 0) & (congratulations = 0) & (win = 0) & (free = 0) & (please = 0) & (service = 0) & (offer = 0) => (spam = 0) OR (spam = 1); [10, 10, 52.63%, 100.00%][1, 9] 
[49, {42, 165, 192, 226, 236, 306, 519, 542, 710}]

Rule 14: 
(sender_known = 1) & (sorry = 1) & (offer = 0) & (great = 0) & (call = 0) => (spam = 0) OR (spam = 1); [9, 9, 47.37%, 100.00%][8, 1] 
[47, 193, 224, 234, 354, 492, 498, 745], {626}]

4 EXPERIMENTAL SET UP AND RESULTS
Matlab language is used for the implementation of the proposed framework. Rose2 software is used for high level results like reduct and decision rules. Naive Bayes and Rough set algorithms have implemented for the Spam filtering task. Extensive tests have been performed with varying numbers of data set sizes. The success rates reach their maximum using all the messages and all the words in training corpus.

5 CONCLUSION AND FUTURE SCOPE
In this dissertation Naive Baye’s has been implemented and test data is giving the desired results. The Naive Baye’s based on Supervised learning technique. We can get association rules from Naive Baye’s but in SMS spam data set we need to find the rules whether an incoming message is spam or not. To implement this is a new mathematical tool rough set has been used. In this dissertation the rudiments of rough set are implemented and high level results i.e reduct and rule induction are obtained by using Rough set tool ROSE2. By the implementation it can be seen that more desired results are obtained by using Rough set theory. In the future improve the structural data the size of Corpus can be implemented by Collecting more SMS and can implement the high level results in the rough sets.

ACKNOWLEDGMENT
The Acknowledgements Authors would like to thank Ms.Richa Arora, Lecture Delhi institute of engineering, Smalkha and Asst. Prof. Ms.Neerja Negi for their supervision during the completion of this work.

REFERENCES


