International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 25219-25227

http://www.ripublication.com

Automated Spam Detection In E-Mail Using SVM

Mrs. Pallavi M S[1] , Mr Vinayak Hegde[2], Ms. Anushadevi H G[3], , Ms. Ambika V[4]
Amrita Vishwa Vidyapeetham, Mysore Campus, #114 Bogadi 2nd Stage Mysore. Karnataka India, palls.ms@gmail.com, Mob No: 8884559752

Abstract

Email Spam is an unsolicited and unwanted email in the mailbox. Spam causes threats to the internet security and it creates traffic in the network. Users spend their precious time to removing all spam emails in their mailbox. Spam email occupies storage space and it consumes network bandwidth and it is very expensive to produce. Detecting these mails and classifying as spam or ham is the major task to the organization. Solution to this problem is filtering the spam emails. Bayesian filter is also a filtering technique. The major issue in Bayesian approach is the performance of filter when word library was very large. In this paper, we introduce the algorithm called Support Vector Machine (SVM) to detect the spam email and send only legitimate email to the users. Support Vector Machine calculates fast and it provides accurate results.

Key words: Support Vector Machine (SVM), email, spam, ham, vocabulary list, classification, Regression.

Introduction

Emails have become an easy and important medium of communication for all internet users. Email spam is the unsolved problems of the today's internet and it is irritated to each individual users and financial damage to companies. Though a study it is found that over 70% of today's business mails are spam [1]. Therefore there have many problems raised with the increasing volumes of spam emails such as wasting of storage space and bandwidth [2], crushing important personal emails and consuming user's precious time to remove all spam mails.

Emails are increasing popularly. Several people found it very easily to distribute unsolicited messages to a number of users at a very low cost. These unwanted bulk messages are called as spam emails [3].Spam emails have become a problem that can affect the usability of electronic mail as a communication. Who sends spam emails are called as spammers. Spammers send spam and try to run their advertisements for companies or products [4].

We are specified some types of spam here. They are: Unsolicited text messages is the costs of sending text messages means spamming can become very expensive. It's easy to trace out the source of text messages and a file. The critical thing here is that users are not trained to click the any preferred links, given the popularity of smart phones that may be prone to malware spread using it. Support Vector Machine is used to detecting these unwanted bulk mails. Phishing scams are the buggers that are designed to look like official emails from financial institutions like eBay and PayPal. This trick is used people into volunteering their usernames and passwords. These are used by the site owners and the scammers to compromise the real accounts Email Spoofing is fraudulent email actually come from a trusted source that is company/organization. It builds the trust in the sufferer, making them more likes to take of the part in whichever scam is included in the mail. Trojan horse mail is the most famous email worm debut in 2000. It is highly succeeded as who won't open an email from a loved one. When it's downloaded, that script attached and damages the local computer and send itself out to everyone the victim knows [5].

We are using Support Vector Machine algorithm to detect the spam email. The existing systems are worked based on the probability. Support Vector Machine is working based on the accuracy. It gives accurate results.

In this paper, we are considering an incoming mail. Initially the email will be read. We are doing pre-processing for that email. The pre-processing step have normalisation, stemming, removing of email contents. We are using vocabulary list to form the 0's and 1's vector representation. Apply the SVM classifier to this vector form and based on the accuracy we are classifying the mail as spam or ham.

The paper is organised as follows: In the introduction part we give a brief introduction on email, spam followed by a literature survey which summarises the related work done using various machine learning methods. In the proposed model we have described the algorithm is SVM (Support Vector Machine) to classify the mail is spam or ham. The experimental results show the percentage of spam or ham mails. We conclude with a conclusion and future work and finally list the acknowledgement and bibliography.

Literature Survey

This section gives a brief literature review of the filtering techniques in spam detection. An ensemble learning and decision tree based approach is proposed to classify the spam emails. Extensive experiments conducted on a public email spam dataset indicate that the proposed algorithm performs the popular classification techniques. Such as C4.5, naive bayes and KNN. A monolingual spam email classification is researched [6]. The popular whitelist, blacklist, greylist, eXpurgate technology, social closeness, adaptive trust management, social interest / disinterest and Bayesian filters approaches that will effectively and accurately block the spam emails and allow only legitimate mails to the user's inbox based on the user's preferences. It is able to determine whether an incoming email message is a spam or not spam [7]. Naive bayes and rough sets methods has a very satisfying performance among the other methods. More research has to be done to enhance the performance

of the naive bayes and artificial immune system or by resolve the feature dependence issue in the naive bayes classifier, or hybrid the immune by rough sets. Finally hybrid systems look to be the most efficient way to generate a successful anti spam filter nowadays [8]. The experiment shows very promising results in the algorithms. Spam recall percentage in the three methods has the less value among the precision and accuracy values. Naive Bayes has a very satisfying performance among the other methods in terms of accuracy.

Proposed System

This chapter describe process of email spam classification. Today many email mail services provide spam filter that are capable of classifying incoming emails into spam and ham. In this we used SVM to build our own spam filter. SVM classify the mails based on accuracy.

Support Vector Machine (SVM) [10] is one of the selective classification methods which are commonly recognized to be more accurate. It is based on the Structural Risk Minimization principle from computational learning theory [11]. It is defined over a vector space where the problem is to find a decision surface that separates the data into two classes. The SVM need both positive and negative training sets which is uncommon for other classification methods. These positive (good) and negative training sets are needed to the SVM to seek for the decision surface that separates the positive data from the negative data in the 'n' dimensional space, so called the hyperplane. The document representatives which are closest to the decision surface are called the support vector [12] [13].

Given a set of training data, which is $\{x_1, y_1\}$ ..... $\{ x_i, y_i\}$ where x_i is the n-tuple of the dataset. y_i belongs to -1, 1 where $y_i = 1$then i belongs to the set A and $y_i = -1$then i belongs to the opposite set.

Support Vector Machine dividing the space by a linear boundary:

$f(x) = \sum^{s}_{k=1} w_k h_k + b$

Where, $\{w_1,.....w_s\}$ are the coefficients of the hyperplane and s denotes the intercept of hyperplane. The output is an example from the test set x_{test} would be $y_{test} = \text{sign}[f(x_{test})]$. So, it belongs to vocabulary list if $f(x_{test})>0$.

Algorithm: SVM-spam-classifier

```
Detect Spam (input_email)
{
   PreprocessMail (input_email)
   {
         Convert_to_lowercase ()
         Word_stemming ()
         Remove_HTML_tags ()
         Normalise_URLs ()
         Normalise_number ()
         Normalise_emailaddress ()
   }
```

```
    Vocabularylist ={ 1 aa,
                      2 ab,
                      3 abii,
                      .
                      .
                      .
                      2000 abc
                }
    for (0 to vocabularylist.size)
    {
for (0 to pre-processed_email.size)
            {
Result=(compare (vocabularylist[i], pre-processed_email[j])
if (result is true)
 vector_y_1++;
else
 vector_y_0++;
}
    }
 TrainData=loadTrainData (train_data.dat)
TestData=loadTestData (test_data.dat)
svmTrain (trainData)
prediction [] =svmTest (testData)
accuracy=findAccuracy (testData, prediction)
spamOrNonSpam = classifyInputMail (input_mail)
return [spamOrNonSpam] }
```

Experimental Results

Pre-processing:

Pre-processing is one of the most important technique in the data mining. This method transforms the raw data into an understandable format. Pre-processing prepares the raw data for further processing of the method. In pre-processing method, initially we are converting all uppercase letters to lowercase, removing all HTML tags from the email. All URLs, email address and numbers are normalized to text format. If words are similar like ‘classify’, ‘classifying’, ‘classified’ and ‘classifies’ are replaced with ‘classify’. This technique is called as word stemming technique. This technique extracts the stem of a modified word. So that the efficiency of any content based spam filter can be improved. Word stemming algorithm has been designed for spam detection. Fig 1 shows a sample data set which is been considered and Fig 2 shows after pre-processing of sample data.

Dear Lucky Winner!!!

CONGRATULATIONS

Dear winner from the Google lottery you have won $500,000.00 USD (Five Thousand U.S Dollars) and 2014 Mercedes-Benz CLA 45 AMG. Contact the claims Manager on Email: (iamscottdaniel@gmail.com), immediately with the following for claims: Name. Address. Telephone. Sex. Date of Birth. Marital status. Occupation. Country.

Figure 1: Before Pre-processing

dear lucki winner congratul dear winner from the googl lottery you have won number number number number number number number number number number usd five hundr thousand us dollar and number number number number mercedesbenz cla number number amg contact the claim manag on email iamscottdanielgmailcom immedi with the follow for claim name address telephon sex dat of birth marit statu occup countr stat

Figure 2: After Pre-processing

Vocabulary list:

The train data and test data are prepared from the vocabulary list as shown in Fig 3. This list is essentially a list of all the words that are present in spam and non spam mails. These words are listed in the list with a unique index number.

```xml
<?xml version="1.0" encoding="UTF-8"?>

    <spam>
        <content>
            <id>1</id>
            <val>not</val>
        </content>
        <content>
            <id>2</id>
            <val>fail</val>
        </content>
        <content>
            <id>3</id>
            <val>ignore</val>
        </content>
        <content>
            <id>4</id>
            <val>profiles</val>
        </content>
        <content>
            <id>5</id>
            <val>claim</val>
        </content>
        <content>
            <id>6</id>
            <val>virus</val>
        </content>
        <content>
            <id>7</id>
            <val>reject</val>
            .
            .
```

Figure 3: Vocabulary List

Training Data Set:

SVM classifier has to be trained with the historical data which is called the training data set as shown in Fig 4. Train data is prepared from vocabulary list. This training data is essentially a group of spam and non spam emails which will be fed to SVM classifier in order to train the classifier to distinguish between the spam and non spam data. Each and every row in a training data set is an email which is spam or non spam. The rows starting with 0 are considered as spam and the row which begins with 1 are considered as non spam mails. Each row represents the combination of x: y for index and word presence. Where x is the index of the word and y represents the word presence.

```
1 1:1 2:1 3:1
1 3:1 4:1 5:1 6:1 7:1
1 1:1 4:1 8:1 9:1
1 1:1 10:1 11:1 12:1
1 1:1 2:1 4:1 7:1 10:1 12:1 13:1 14:1
.
.
.
0 15:1 16:1 17:1 18:1 19:1
0 16:1 20:1 21:1 22:1 23:1 24:1 25:1 26:1
0 16:1 27:1 28:1 29:1 30:1 31:1
0 22:1 23:1 24:1 26:1 32:1 33:1 34:1
0 16:1 18:1 22:1 28:1 35:1 36:1 37:1 38:1 39:1
0 115:1 201:1 141:1 202:1 146:1 203:1 204:1 205:1 206:1
.
.
.
```

Figure 4: Training Data Set

Test Data Set

SVM classifier is tested by the test data as shown in Fig 5. Test data is prepared from the vocabulary list. This test data is essentially a group of spam and non spam mails which will be fed to SVM classifier in order to test the classifier to distinguish between the spam and non spam data. Each and every row in a test data set is a mail which is spam or non spam. The rows starting with 0 are considered as spam and the row which begins with 1 are considered as non spam emails. Each row represents the combination of x: y for index and occurrences. Where x is the index of the word and y represents the number of occurrences.

```
1 0 115:1 201:4 141:2 202:2 146:1 203:1 .......
```

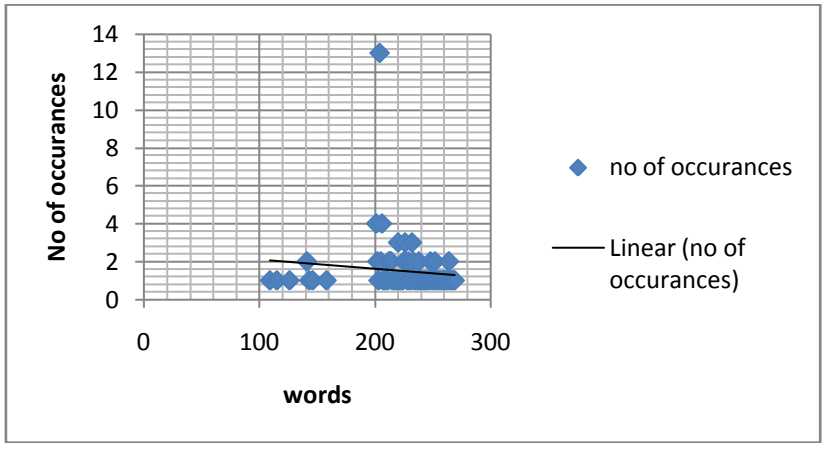
Figure 5: Test data

Classification/Prediction

A hyper plane / regression line is plotted as shown in Graph 1 and margin of the line is determined by the traning data as shown in Fig 6. Based on the margin detected the email is classified as either spam/ham.

index no	no of occurances
109	1
115	1
126	1
141	2
143	1
146	1
158	1
201	4
202	2
203	1
204	13
205	2
206	4
207	1
208	1
209	1
210	1
211	1
212	2
213	2
214	2
215	1
216	1
217	1
218	1
219	1
220	3
221	1
222	1
223	1
224	1
225	2

Figure 6: Test Data ready for plotting hyper-plane / regression line.



Graph 1: Hyper-plane/ Regression Line.

Conclusion

Spam emails are the big problem for the internet data. This paper explores collaborative approach to solve this problem. Support vector machine classifier provides an efficient mechanism for spam filtering. SVM-spam-classifier outperformed compared to other classifiers for the same data set. It can be concluded that SVM-spam-classifier are capable of spam.

References

[1] Hassan Najadat and Ismail Hmeidi, "Web Spam Detection Using Machine Learning in Specific Domain Features", Journal of Information Assurance and Security 3 (2008) 220-229.

[2] Kalpana and B. Venkateswarlu, "Cosdes: A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme", IJCSET, October 2012, Vol 2, Issue 10, 1447-1451.

[3] Sayyed Saissen and Sayeed Yasin, "A Novel Method for Collaborative Spam E-Mails", Intenational Journal of Modern Engineering Research (IJMER), Vol. 3, Issue. 3, May – June 2013 pp-1825-1827.

[4] R. Kholghi, E. Eidkhani, S.A. Mortazavi, M. Hajyvahabzadeh and A. Nemaney Pour, "A New Method to Stop Spam Emails in Sender Side", ACEEE Int. J. On Network Security, Vol. 02, No. 04, Oct 2011.

[5] http://www.ranker.com/list/types-of-spam-e-mails/kel-varnsen.

[6] Lei SHI, Qiang WANG, Xinming MA, Mei WENG and Hongbo QIAO, "Spam Email Classification Using Decision Tree Ensemble", Journal of Computational Information Systems 8: 3 (2012) 949-956.

[7] Amandeep Kaur and Malti Sarangal, "A Hybrid approach for enhancing the capability of Spam Filter", International Journal of Computer Applications Technology and Research, Volume 2- Issue 6, 759 – 762, 2013.

[8] W. A. Awad and S. M. ELseuofi, "Machine Learning Methods For Spam Email Classification", International Journal of Computer Science and Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[9] SS Sivatha Sindhu, S Geetha, A Kannan, Decision tree based light weight intrusion detection using a wrapper approach. Expert Syst. Appl **39**, 129–141 (2012).

[10] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan, "A Review of Machine Learning Algorithms for Text- Documents Classification", journal of advances in information technology, Vol.1, NO. 1, Feb 2010.

[11] Vladimir N. Vapnik, "The Nature of Statistical Learning Theory" , Springer, NewYork. 1995.

[12] Wen Zhang a,*, Taketoshi Yoshida a, Xijin Tang, "Text classification based on multi-word with support vector machine", Knowledge-Based Systems 21 (2008) 879–886.

[13] J.W. Han, M. Kamber, Data Mining Concepts and Techniques, second ed.,Morgan Kaufmann Publishers, 2006.