Security in Email Systems using Spam Filtering

(Support Vector Machine Concept)

Pravin Khedkar[1], Rohit Reshi[2]
[1, 2] Computer Department,
[1, 2] AISSMS's IOIT

Abstract- *The most dangerous online threat now-a-days is the threat in the mailing system. The spam are the threat in the mailing system which are any unwanted and harmful mail for the security purpose. The spam mails should separate from the rest of mails which are useful to the users. This paper implements spam filtering techniques using Support Vector Machine (SVM). Techniques to separate spam mails are word based. Content based, machine learning based and hybrid. The machine learning techniques are most popular because of high accuracy and mathematical support. SVM is used mostly for machine learning based technique because of its ability to handle data with large attributes. SVM training time minimization is the global minimization problem. During training, entire dataset is used to get the final output, so there is a lot of room for parallelization*

Keywords- Spam filtering techniques • Word based • Content based • Machine learning based • Support Vector Machine (SVM)

I. INTRODUCTION

Material handling is one of the important phase of In modern day communication tools, e-mail system is mostly used. E-mail became boon for business because of its wide availability-mail is fastest way of communication as there is no need to wait for response. The main danger for the e-mail is spam mail. Unwanted mail is also known as spam mail. Mails which are sent in bulk are spam mails. Phishing websites, malicious attachments are sent by spam e-mails. Spam e-mails also include malicious scripts and executable attachments. The threat and the major problem like this of getting unwanted and malicious programs motivated us to build the system which separates spam but also blocks the accounts sending it.

II. GOALS AND OBJECTIVES

- With increasing security measures in network services, remote exploitation is getting harder. Therefore efficient filtering methods for spam messages are needed.
- In this project, we introduce a more proactive approach that allows us to detect valid and spam mail.
- The main aim is to design and develop a spam detecting system for emails using classification algorithm i.e. SVM

III. PROPOSED TECHNIQUE

A. Support Vector Machine

Support Vector Machine (SVM) based approaches have persistently gained popularity in terms of their application for text classification and machine learning. Classification in SVM based approaches is founded on the notion of hyper planes. The hyper planes act as class segregators in common binary classification, such as spam or ham in the context of spam filtering. SVM training is a computationally intensive process. Many SVM formulations, solvers and architectures for increasing SVM potential have been explored and proposed including distributed and parallel computing techniques. SVM decomposition is another widespread technique for improving the performance in SVM training. Decomposition approaches work on the basis of identifying a small number of optimization variables and tackling a set of fixed size problems. Another widespread and effective practice is to split the training data into smaller fragments and use a number of SVM's to process the individual data chunks. This in turn reduces overall training time. Various forms of summarizations and aggregations are then performed to process the final set of global support vectors. Numerous forms of decomposition which are based on a data splitting strategy approach can suffer from issues including convergence and accuracy. Challenges related to chunk aliasing as well as outlier accumulation tend to intensify problems in a distributed SVM context. Adopting a training data set splitting strategy commonly amplifies issues related to data imbalance and data distribution instability. In this paper, we present an ontology assisted, parallel scheme for scalable SVM training. This work supplements current approaches by focusing on a number of aspects. We prototype a parallel SVM, building on the Sequential Minimal Optimization (SMO) algorithm.

Support vector machine is techniques used in text classification. In this pattern recognition and data analysis is done. In machine learning the training sample is a set of

vectors of n attributes. Assume that we are in a hyperspace of n dimensions, and that the training sample is a set of points in the hyper-space.

Let us consider the simple case of just two classes (as it is the case of spam problem).The classification using Support Vector Machine look for the hyper plane able to separate the points of the first class from those of the second class. Distance between the hyper plane and points of each class, is kept maximum for good separation.

This insures the existence of hyper plane that separates the two classes. One interesting feature is that to find the appropriate plane, SVM method explore just the nearest points.

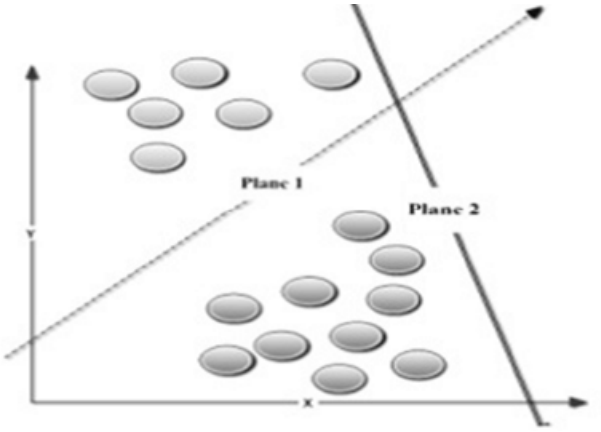

Figure 1. Hyper-plane that separate two classes

IV. KEY TECHNOLOGY

A. Support Vector Machine

In 1995, Vapnik and his colleagues proposed a novel method called support vectors machine (SVM) to perform pattern categorization. Recently SVM has been successfully applied to many domains including hand written digit recognition, text categorization, speech and recogmtlon, bioinformatics, and etc. SVM is a family of machine learning theory. Assuming a given set of training samples {(x1,y1), (x2 ,y2),….,(xn ,yn)} where xi ϵ {-1,+1}} ,i=l, ... ,n, Rd is d-dimensional Euclidean space, Yi the two types of training samples class target.

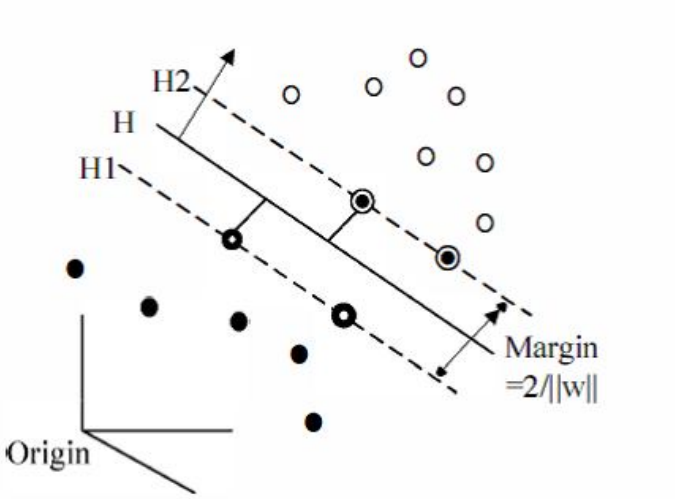

Figure 2. The optimal separating hyperplane under linear separability

In Fig. 2, Hollow-point and solid-point are two training samples, while H is categorization plan, HI and H2 planes that are respectively through proximate samples away the various categorization plan and parallels to it. The distance between HI and H2 is called interval (It is Margin). If all the samples in training corpus can be correctly plotted out by a hyperplane and the distance of proximate vectors away plan H, we call it different-vectors, achieves the largest value, this plan is deemed to the most optimal categorization hyperplane (Plane H in Fig. I). Its equation is $WtX + b = 0$, where w is the normal of categorization plan in that vector w is. We call this different-vectors support vector, as the point including double circle in Fig. 2. A vector group can only determine a hyperplane.

B. Mail-header feature Pretreatment

As we all know, in the entire process mail transfer chiefly depends on five SMTP core orders: HELO, MAIL FROM, RCPT TO, DATA and QUIT. Mail senders mark his identity using HELO. MAIL FROM indicates the beginning mail transmission, meaning "I have to send mail from a person". The address following this order is the so-called "envelope address". Of course, this envelope address is not necessarily full by mail senders. RCPT TO and MAIL FROM are complementary. They can designate mail recipients. Through a number of orders RCPT TO a mail can be sent to many recipients. DATA instruct the actual position of the mail content transmission. QUIT can terminate this connection. It shows that seven fields from the mail header are likely to contain fake data. These fields are from field to field, reply-to field, delivered-to field, return-path field, and data field.
In these fields, the following situations may arise.

- This field is NULL
- This field is blank;
- The user name of this field is blank
- The domain of this field is blank
- The address format of this field is mistake
- The domain of this field is spoofing;
- The address of this field has two @;
- The address of this field does not have @;
- There are not domain and user name, just only @;
- The data field is overdue;
- The receive field contains too much paths.

V. IMPLEMENTATION AND TEST PERFORMANCE

Mail feature categorization system is made up of the training module and categorization module. Training module quantifies the training samples to produce the eigenvectors indicating mail header information, and then extract the 106 eigenvectors from feature set by feature extraction algorithm to concentrate the text vectors. Finally we can get SVM classifier by the training algorithm, shown here as a MODEL file. Categorization module quantifies the test document for the first place to generate eigenvectors as same as that in the MODEL file. Then we classify the test document through the trained classifier (the MODEL file). Specific process is as in the Fig. 3 below.

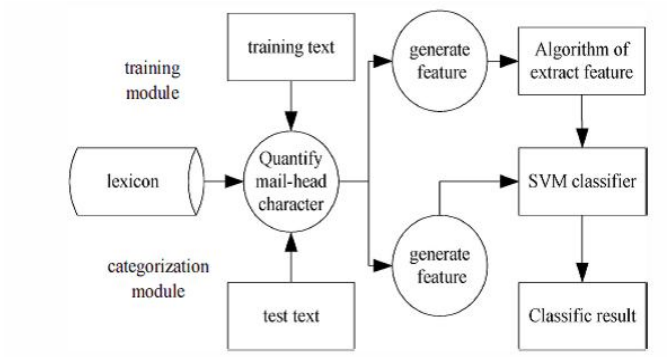

Figure 3. Features chart of mail-head Categorization System

To test the performance of this system we require a public corpus, usually including training set and test set. To remove the influence brought by the contents of the corpus to algorithm it is general that the data of corpus is from the real application. Currently, the most common corpuses are Reuters English corpus, the International Text retrieval Conference corpus, the People's Daily Chinese corpus, and etc. Building corpus in text processing research is a very critical issue too. Here, we test the SVM categorization system using PU I corpus. The legitimate mail used in PUI is from the two parts.

One part is the legitimate mail sent by senders who is out of address table. Another part is the first five letters sent by senders in address table. The senders will be joined the table if the first five letters he or she sent is legitimate mail. A total of 1,099 are used in PU I corpus, including 481 spam and 618 non-spam. Here we indicate the test results using the accuracy that is the ratio between the mail header feature information classified correctly and the total one. In this paper SVM categorization system is sponsored by our project-Research on Spam Detection and Control Key Technology. We construct it through the lib of LIBSVM, using a radial basis function (RBF) as core function for training the categorization system.

We choose the most common method in spam filtering system, Naive Bayes, to compare with this technology based on SVM. At the same time to evaluate this categorization system's performance on the use of different number of test corpus, we tested this system under different number of mail information from 5,000 to 25,000. The test results as shown in table 1.

Table 1. Under Different Test Items SVM and BAYES Accurate Ratio

Number	SVM(RBF)	Bayes
5000	92.15	98.74
8000	93.96	97.31
10000	94.32	98.60
15000	93.06	93.32
18000	95.80	90.01
20000	96.91	87.69
25000	98.11	82.87

From table 1 we can see that SVM categorization accuracy is not very clear under the different feature number from 5000 to 8000. However following the increasing number of features, SVM categorization accuracy is getting higher and higher. Our scale of experiment is still not big enough, not analyzing its performance in a variety of types distributing situation, so it is not enough to explain the merits of SVM comparing with Bayes. However, it can be considered that the SVM is a categorization method with outstanding performance.

VI. THE ADVANTAGES OF THE SVM TECHNIQUE CAN BE SUMMARIZED AS

- By introducing the kernel, SVMs gain flexibility in the choice of the form of the threshold separating different companies.
- Since the kernel introduced by the SVM, a non-linear transformation assumes nothing about the form of the data transformation, which makes data necessarily separable.
- SVMs provide a good out-of-sample generalization.
- A common disadvantage of techniques such as SVMs is that is doesn't has the transparency to show the results. Being the dimensions too high, the score of different or all companies can't be represented by simple parametric functions of the financial ratios. The financial ratios don't have the constant weights. Thus the variable score is generated from the marginal contribution of each financial ratio.

VII. CONCLUSION

- Developed a mailing system capable of isolating the spam mails and containing them separately in spam folder. Also it includes the blocking of particular account for sending spams emails in bulk.
- The email system which will take input as the email received by its users , the received email is then validated as 'spam' or 'ham' by comparing it to the keywords used as a spam mail using the algorithms.
- As the spam emails can be send or are sent in a bulk in normal email system, in our system the account sending spam mail in bulk will be blocked.

VIII. FUTURE SCOPE

SVM is used mostly for machine learning based technique because of its ability to handle data with large attributes, there are also some hurdles in the training process of SVM, that can't be given as input, both of these problem should be solved by implementing the training algorithm on map reduce (Hadoop) framework which gives up to 6 times speedup than sequential algorithm.

- **Map Reduce**

Training the SVM set completely is a bit difficult, so the input data is divided into various sub-sets of data and is individually being worked on. Generally Sequential Minimal Optimization (SMO) is used to do that. The only problem with SMO is that it can't handle large data sets. So this problem could be eradicated by using the Map Reduce framework. A Map Reduce framework usually splits the training data-set to various different independent chunks of data sets which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. The Map Reduce framework is most popularly used in the research fields like Mars, Phoenix, and Hadoop etc.

REFERENCES

[1] Amol G. Kakade, Prashant K. Kharat," Spam filtering techniques and Map Reduce with SVM: A study", 2014 Asia-Pacific Conference on Computer Aided System Engineering (APCASE).

[2] Saadat Nazirova," Survey on Spam Filtering Techniques", Communications and Network, 2011, 3, 153-160 doi:10.4236/cn.2011.33019 Published Online August 2011.

[3] Rekha, Sandeep Negi ," A Review on Different Spam Detection Approaches", International Journal of Engineering Trends and Technology (IJETT) – Volume 11 Number 6 - May 2014.

[4] J. Vijaya Chandra, Dr. Narasimham Challa,, Dr. Sai Kiran Pasupuleti, "A Practical Approach to E-mail Spam Filters to Protect Data from Advanced Persistent Threat", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT].

[5] Tarjani Vyas, 2Payal Prajapati, & 3Somil Gadhwal," A Survey and Evaluation of Supervised Machine Learning Techniques for Spam E-Mail Filtering",

[6] Wanqing You, Kai Qian, Dan Lo, Prabir Bhattacharya, Minzhe Guo, Ying Qian," Web Service-enabled Spam Filtering with Naïve Bayes Classification", 2015 IEEE First International Conference on Big Data Computing Service and Applications.

[7] Anirudh Harisinghaney, Arnan Dixit, Saurabh Gupta, Anuja Arora," Text and Image Based Spam Email Classification using KNN, NaIve Bayes and Reverse DBSCAN Algorithm", 2014 International Conference on Reliability, Optimization and Information Technology ICROIT 2014, India, Feb 6-8 2014.

[8] Godwin Caruana1, Maozhen Li1,3 and Man Qi2," A MapReduce based Parallel SVM for Large Scale Spam Filtering", 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)

[9] Amol G. Kakade1, Prashant K. Kharat2, Anil Kumar Gupta," Survey of Spam Filtering Techniques and Tools,

and MapReduce with SVM ",IJCSMC, Vol. 2, Issue. 11, November 2013, pg.91 – 98 .

[10] Salwa Adriana Saab, Nicholas Mitri, Mariette Awad," Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering", 17th IEEE Mediterranean Electrotechnical Conference, Beirut, Lebanon, 13-16 April 2014.