A Survey on Spam Filtering Methods and Mapreduce with SVM

Dipika Somvanshi[1], Prof. Kanchan Doke[2]

[1] ME(Comp.) Student of Bharati Vidyapeeth College of Engineering, Kharghar, Navi Mumbai ,India
[2]Proffesor, Computer Dept., Bharati Vidyapeeth College of Engineering, Kharghar, India

---***---

Abstract - *Spam is any unwanted and harmful mail send to massive recipients in bulk quantity. Spam can be harmful as it may contain malware & links to phishing websites. So Separation of spam from normal mails in separate folder is essential. Techniques to separate spam mails are word based, content based, machine learning based and hybrid. Machine learning techniques are most popular because of high accuracy and mathematical support. This paper surveys different spam filtering techniques. SVM is the popular machine learning techniques in spam filtering because it can handle data with large number of attributes.*

SVM requires more time to train the data and for training it can't work with large a dataset, these drawbacks can be minimized by introducing MapReduce framework for SVM. MapReduce framework can work in parallel with input dataset file chunks to train SVM for time reduction. This paper aims at surveying of few such spam filtering techniques and scope to introduce MapReduce with SVM.

***Key Words*: Spam Filtering; Machine Learning Techniques; Naïve Bays; KNN; Decision Trees; SVM; Mapreduce**

1. INTRODUCTION

The email system is one of the most used communication tools. Email is a quick means of communication because one has not to wait for the response and it is straightforward way to stay in touch with the all. One major threat to an email system is spam e-mail. The spam e-mail is nothing but the unwanted mail send in bulk quantity by spammers group for their advantage. Spammers are group of people intended to spread malicious content, advertise content, Links to phishing websites through email. Spam Emails causes overloading of server bandwidth, storage, cost, time for separation of spam emails from ham E-mails. According to the SMX email security provider, the live spam percentage is about 79.5%. The average size of spam is 16 KB[1]. So classification of Emails in spam & ham is most important issue.

For the separation of such spams from important mails, spam filtering is important. Various spam filtering techniques exist in literature survey. Spam filtering techniques are classified as Machine Learning based, Content based, List based, Hybrid. Amongst them Machine learning techniques give more accurate results due to their mathematical background. Machine learning techniques works with data mining algorithms and gives more satisfied results. For spam filtering filters are trained with algorithms for sample data set of emails & then tested for new sample of emails. Machine learning based spam classification algorithms are SVM, Naïve Bays, KNN, Decision Trees, etc. Amongst these, Naïve Bayesian classification and Support Vector Machine are most used and appreciated by researchers. Also, number of freeware and paid tools are available for spam filtering, they also makes use of these techniques. Support Vector Machines (SVM) can be applied efficiently in spam filtering. SVM works with Kernel function and gives most satisfied results in spam filtering. SVM works best with small set of data input. But it's performance degrades with increase in size of dataset. It requires large time to train filter. So this issue needs to be addressed.

MapReduce can be effectively used for training of large dataset input. It can process large data within less time. So in this proposed system we have used MapReduce with SVM to classify large set of emails into spam and ham. MapReduce with SVM gives more speedup than traditional SVM algorithm.

2. LITERATURE REVIEW OF MACHINE LEARNING TECHNIQUES

A. Clustering:

Clustering is used for separation of objects into relative classes called clusters. It classifies object or observations in such a way that objects in a group are more similarto each other than tothose in other group.[2]

***K- Nearest Neighbor**:*

It is one of the simplest machine learning algorithms. KNN works with 'characteristics vector'. The characteristic vectors are measure of similarities among all messages. In this algorithm new incoming email is classified on the basis of distance of that mail from both classes. Distance can be calculated by Euclidian distance. KNN works as follows. It consists of two phases.

Training phase: Storing the feature vectors & class labels of training emails.

Classification Phase: *k* is a user-defined constant, and an unlabeled email is classified by assigning the label which is most frequent among the *k* training email samples nearest to new incoming email.

B. Bayesian Classification:

Bayesian classification is based on Bays theorem. Bayesian classifiers can predict class membership probabilities. Membership probabilities are the measure of probability that given sample belongs to a particular class.

In case of spam classification, probability of words is calculated for spam and ham emails. Probability that particular word is occurred in spam emails & ham emails is calculated. Depending on that a training dictionary is generated. When new email is arrived, its words probability is calculated & if spam words probability is more than ham probability then it is classified as spam email else ham.

Naive Bayes classifier calculates word probability as follows:

$$\Pr\left(\frac{S}{W}\right) = \frac{\Pr\left(\frac{W}{S}\right)}{\Pr\left(\frac{W}{S}\right)+\Pr\left(\frac{W}{H}\right)}$$ ----------Eq. No. (1)

Where,

Pr(S/W) = probability that a message is a spam.

Pr (W/S) = probability that the specified word appears in aspam message.

Pr (W/H) = probability that specified word appears in hammessage.

It has complexity of O(n) (where n is total number of observations).

Drawback of this classifier is that it uses strong independence assumption between features but in real world problem it may not be applicable.[2]

C. Decision Trees:

Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node has two branches spam or ham. Topmost node is root node corresponds to most predictor. ID3 is a non-incremental algorithm used to build a decision tree from a fixed set of observations (dataset).The resulting tree is used to classify test observation. Each observation is represented by its features or attributes and a class to which it belongs. ID3 uses information gain measure to select decision node. Information gain indicates the ability of a given attribute to separate training examples into classes. Higher the information gain, higher is the ability of the attribute to separate training observation. Information gain uses entropy as a measure to calculate the amount of uncertainty in dataset.[2]

It builds fastest and short tree and considers attributes that are enough to classify data. But it suffers from over-fitting problem if training data is small.[2]

D. Support Vector Machine(SVM):

SVM is a of machine learning algorithmbased on statistics learning theory. SVM is a kernelbased technique widely used for classification, regressionand outlier detection. The main reason of itsincreasing importance is its ability to cast nonlinearclassification problem as a quadratic problem (QP). Nowadays there is a development of special purposealgorithm for solving QP. Sequential minimaloptimization (SMO) has been used for faster training ofSVM model.

Advantages of SMO are that it works effectively in highdimensional space. It also gives good results whennumbers of dimensions are greater than the number ofobservations. Also it is memory efficient.

Disadvantage of SMO is that it can’t handle large data set & its time consuming for large dataset as input.

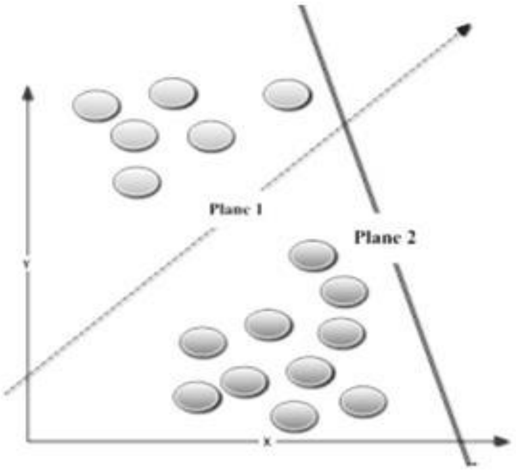

Fig-1Hyper plane that separate the two classes

SVM Working:

1. Every mail instance is treated as a single point with *n* dimensions in hyperspace.
2. The distance between the hyperplanes and points of each class, is kept maximum, for good separation. Here in fig-1, Plane1 is good classifier and Plane2 doesn't classify all instances.
3. It may also happen that, we can't find good separator hyperplane (Plane 1) as in fig.1. In suchcase, hyper space is called as non-linearly separable.
4. To obtain linear separation in the non-linearly separable hyperspace, it is extended to more dimensions.[1]

3. PROPOSED SYSTEM

Proposed system contains following modules:

3.1 Admin Module

The admin has the complete right over the system. The admin can view List of Valid and Invalid i.e spam mail.

- **Add training E-mail:**This Module will facilitate Admin to specify sample e-mail content containing valid or spam email. This email will be further analyzed to store some training parameters as valid or spam email.
- **Training Module:**In this Module, Adminwill specify sample training data for valid and spam mails and divide training set of emails into 2 random set of emails on which training will be performed and result will be displayed in list as valid email list or spam email list. Training is provided based on MapReduce with SVM model.

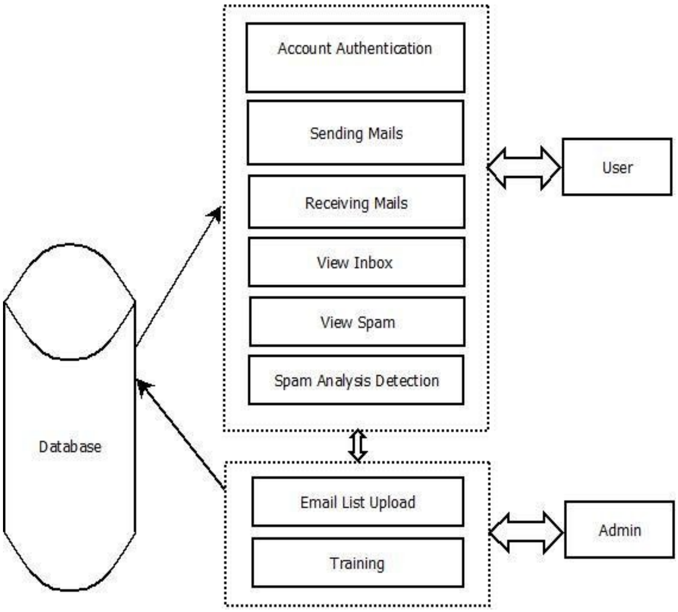

Fig.2 Architecture of proposed system

3.2 User Module

The user can be defined as an individual accessing the application. User has the features of sending and receiving the emails within the domain of app (local domain) and viewing Inbox and Spam Emails

- **Account authentication:**Through this feature, a user is authenticated into the system. The login credentials of the user must match the information stored in the application. On successful provisioning of the credentials, a user is logged in the system
- **Sending mails:**The user of the application has the right to send across a regular text email to a desired receiver of the choice.
- **Receiving Mails:**The user of the application has the right to receive emails from other senders in the application.
- **View Inbox:**The User of the applicationhas right to view Inbox.
- **View Spam:**The User of the applicationhas right to view Spam.
- **Spam Analysis Detection:** Whenever a user sends across an email, it passes through the Spam Analysis detection stage which is a vital module

used for spamDetection based on which the system will generate a list containing Valid and Spam mails from current user.

3.3 Spam Analysis based on MapReduce with SVM model

In this paper, a M/R based distributed SVM algorithm for scalable spam filter training, designated MRSMO, is presented. By distributing and processing subsets of the training data across multiple participating computing nodes, the distributed SVM reduces spam filter training time significantly. This approach minimizes memory requirements drastically to store the matrix of input mail instances. With MapReduce large input E-mail file for training SVM is spited into small size data chunks. Input data chunks are treated as Key (K) and Value (V). In case of spam filtering, Label(Spam/Ham), Weight(highly weighted top 10 keywords) Each map task works with one single data chunk, so numbers of data chunks are normally equal to the number of map tasks. Fig. 2 shows splitting of input file and working of MapReduce paradigm. For each E-mail, each Map task run independently with serial SMO on their respective training set. Output produced by map tasks is passed through shufflers, combiners, sorters etc to group same Keys (K) at near to each other. MapReduce gives output in the form of {key, value} pair. The Reduce task has {Label, Value} pairs as an input, generated by each Map task. Then it combines the result of all Map tasks to get final output.[1,2]

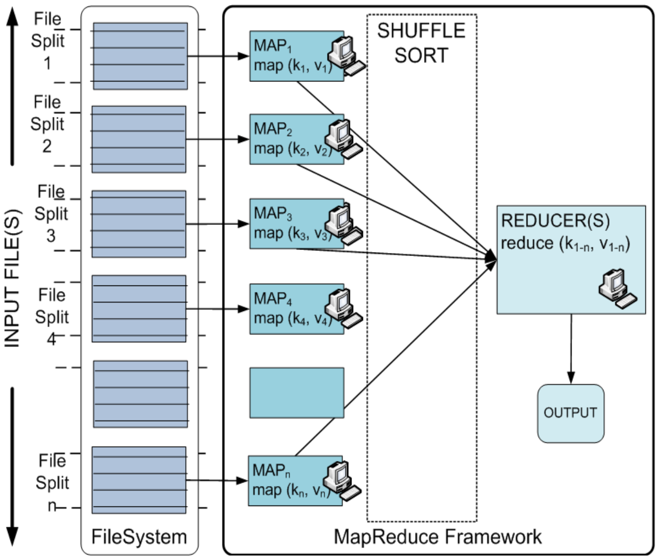

Fig.3 MapReduce Architecture

4. COMPARISON OF CLASSIFICATION ALGORITHMS WITH MAPREDUCE

Classification Algorithm	Traditional Model (Without MapReduce)	With MapReduce Model
SVM	1. Low Scalbility 2. More time required to train the filter 3. Works best with small input dataset.	1. High Scalability 2. Speedup training process 6 times as compared to traditional approach. 3. Works best with large datasets also
Bayesian Classification	1. Strong independence assumptions between the features limit performance accuracy 2. Low Scalability	1. Higher spam detection accuracy & speed. 2. High Scalability
Clustering	1. Computation cost is very high because need to calculate distance of each query instance to all training sample emails 2. Can't handle large dataset	1. Computation cost is very low as all instances tested in parallel way efficiently 2. High Scalability using datasets up to 1 million instances.[3]

5. CONCLUSION

Traditional SVM without MapReduce gives low scalability and require high computation time as well as cost. MapReduce framework is designed to provide large scalability, Speedup and high accuracy. In similar way, Bayesian and Clustering algorithms performance increases with MapReduce model. But, Decision tree shows limitations while implementing MapReduce due to its irregular nature. Its performance depends on large number of attributes, shows small performance gain. So, We conclude that SVM with MapReduce framework gives most satisfied results in spam filtering.

6. ACKNOWLEDGEMENT

I take this opportunity to express my intense gratitude and deep regards to my Guide Prof. Kanchan Doke for her guidance, monitoring and constant encouragement throughout this paper.I would like to express deepest appreciation towards all staff members of BVCOE, Navi Mumbai.

7. REFERENCES

[1]“Spam Filtering Techniques & Map Reduce with SVM”, Amol G. Kakade, Prashant K. Kharat, Anil Kumar Gupta, Tarun Batra, Department of Information Technology,2014 Asia-Pacific Conference on Computer Aided System Engineering (APCASE), 978-1-4799-4568-9/14/$31.00 ©2014 IEEE.

[2] ”A Survey and Evaluation of Supervised Machine Learning Techniques for Spam E-Mail Filtering”, Tarjani Vyas, Payal Prajapati, & Somil Gadhwal, Institute Of Technology, IEEE international conference on Electronics, Computers & Communication Technologies, 978-1-4799-608S-9/1S/$31.00©2015 IEEE

[3] ”A MapReduce-based k-Nearest Neighbor Approach for Big Data Classification” , Jesus Maillo, Issac Tringuero, Francisco Herrera, IEEE Trustcom/BigData/ISPA978-1-4673-7952-6/15 $31.00 © 2015 IEEE

[4] “A MapReduce Implementation of C4.5 Decision Tree Algorithm”, Wei Dai, Wei Ji, International Journal of Database theory and Application,Vol.7, No.1 (2014), pp.49-60

[5] “ Survey on Spam Filtering Techniques”, Saadat Nazirova , *Institute of Information Technology of Azerbaijan National Academy of Sciences, Communications and Network*, 2011, 3, 153-160

[6] “A TAXONOMY OF EMAIL SPAM FILTERS”, HASAN SHOJAA ALKAHTANI, PAUL GARDNER-STEPHEN, AND ROBERT GOODWIN, Computer Science Department, College of Computer Science and Information Technology, King Faisal University, Saudi Arabia

[7] “Origin (Dynamic Blacklisting) Based Spammer Detection and Spam Mail Filtering Approach”, Nikhil Aggrawal, Shailendra Singh, Dept. of Computer Science, IEEE international paper on computer & Networks, ISBN: 978-1-4673-9379-9 ©2016 IEEE

[8] “EVALUATION OF DECEPTIVE MAILS USING FILTERING & WEKA”, Sujeet More, Ravi Kalkundri, IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS’15, 978-1-4799-6818-3/15/$31.00 © 2015 IEEE

[9] “Spam Filtering by Semantics-based Text

Classification “,Wei Hu, Jinglong Du, and Yongkang Xing, 8th International Conference on Advanced Computational Intelligence, Chiang Mai, Thailand; February 14-16, 2016, 978-1-4673-7782-9/16/$31.00 ©2016 IEEE

[10] “Spam Classification Based on Supervised Learning by Machine Learning Techniques”, Ms. D. Karthika Renuka, Dr. T. Hansapriya,Mr. M. Raja Chakravarthi, Ms. P. Lakshmi Surya, IEEE international paper on Data mining, 978-1-61284-764-1/11/$26.00 ©2011 IEEE