International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

Comparative analysis of classifiers for header based emails classification using supervised learning

Priti Kulkarni[1], Dr. Haridas Acharya[2]

1Assistant Professor, Symbiosis Institute of Computer Studies and Research, Symbiosis International University, Pune-16, Maharashtra, India
2 Professors, Allana Institute of Management Sciences, Pune-16, Maharashtra, India.

Abstract - *Emails are used as primary communication tool in business. They are preferred as fast means of communication. In the business domain, they facilitate information-gathering and communication function. The volume of emails receiving into an inbox varies from tens for regular users to tens of thousands for an enterprise.*

Inbox may consist of some unwanted emails, called spam emails. This results into unnecessary consumption of both band width and inbox space. It is important to classify unsolicited emails receiving in the inbox, so that they can be filtered out at the right stage.

This paper compares the various existing techniques for spam classification using email header fields. Paper also presents a discussion on challenges for spam filtering.

Key Words: *spam, ham, filtering, email classification, header*

1. INTRODUCTION

The Internet is becoming an integral part of our everyday life. Due to speed and lower cost email is treated a powerful tool for information exchange. This has changed the way business works. However rapidly increasing volume of emails received is a matter of concern, especially when spam mails contribute a lot to congestion and waste of Internet resources.

According to a research survey, by GFI software carried out in 2014, 68.5% employee in company said that they had spam related disruption to affecting their business operations. So there is a need to automate the process of blocking of spam email messages.

This paper is organized as follows. To begin with the concept of spam is introduced. The second section describes different machine learning techniques used for email classification. Third section discusses the challenges for antispam techniques. The paper intends to cover comparison of spam filtering techniques using email header features.

1.1 SPAM - (Unsolicited Bulk Email)

Spam Email is an email sent by an unknown sender, generally sent as a part of bulk email to large groups with commercial nature. Spam emails are also identified as U.B.E. (Unsolicited Bulk Email). Jon Postel, an Internet pioneer, in 1975 has addressed the problem of junk mail in Requests for Comments (RFC) 706. Later on in April 1994 the word spam became popular.

The TREC Spam Track gave similar definition: Spam is "unsolicited, unwanted email sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user" [1]. A varient of this definition can be found in [2]. Radicati Group survey -2013- 17 predicted increases in email account from 3.9billion email accounts in 2013 to 4.9billion email accounts till the year 2017. [3] According to another survey conducted in 2014 maximum spam comes from China, US and South Korea [4], and it estimates that American firms and consumers experienced cost of $20 billion annually. According to survey 14.5 billion are spam messages globally per day. 52% participants stated spam is major problem. 31.7% spam are adult related subjects, 26.5% are financial matter [5]. Similar facts have been reported in [6]

Generally, spammers send spam emails for purposes like financial gain, to reduce competitors productivity, and for advertising.

1.2 Email Structure

RFC 822 has defined a structure for Standard for ARPA Internet Text Messages and later the RFC 2822 modified syntax of RFC 822. RFC 2045 specified format of Multipurpose Internet Mail Extension header. Email message is divided into two parts: body and header. Header of email stored additional information about the message. Each email includes structured and standard data fields such as primary

recipient (To), copied recipient (cc), blocked receiptant (BCC), subject line, sender and date. The body text of the email is unstructured and body contains basic contents of an email message containing text, image, and even video.

2. LITERATURE REVIEW

It is difficult for antispamming technique to filter spam, mainly because; spammers are using different techniques to bypass the email server.

According to literature emails are classified as spam and nonspam by two methods namely content filtering method and header base method.

Authors of [7] have proposed SVM as the method to classify emails into spam and non spam. Whereas authors of [8] Have selected few header fields "To", "Cc", "From", "Date" and "Subject" to characterize emails. Youn et al [9] have used body of email as the base for classification, and have compared Neural Network, SVM, Naive Bayesian and J48 classifiers. They have concluded that NB and J48 shows better accuracy than SVM and NN. Similar study has been presented by Rafiqul Islam et al [10].

M. Basavaraju et al [11] proposed the text based clustering method for spam detection. They used BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) method to clustering the documents.
Jefferson Provost,[12] used subject and body as features and compared Naïve Bayes algorithm with RIPPER rule learning and showed that Naïve Bayes (87%) achieve better accuracy over RIPPER (78%). Liu Pei-yu et al (2009) [13] suggested a method of improved Bayesian algorithm for filtering spam. KNN algorithm SVM, decision tree, and improved Bayesian algorithm are used for classifying texts. Improved naive Bayesian algorithm is a combination of Bayesian algorithm with boosting method, developed to reduce the rate of misjudgment and improve the accuracy of classification. Zhang et al. [14] conducted an experiment using Body of message and showed that the performances of a handful of machine learning algorithms are not satisfactory with text features. Sheu classified emails by extracting features from email header into four categories as sexual, finance, job hunting, marketing, and advertising are total categories. [15] He has used sender's field, subject, sending date and size of email as features. Wang and Chena [16] used mail user agent, message-id, sender and receiver addresses as features. They applied statistical analysis of the header session message and results demonstrated up to 92.5% accuracy to filter junk emails. Content based filter works on text data and mainly focused on spam terms use for classification. [17]authors had address the problem of filter evasion. This is problem of altering text of messages so that message will not filtered by classifiers.

Image spam is another area of importance. Spammers use text of spam messages in readable images. So instead of passing text as body of message, images are used for passing spam. These images are attached with body of emails. Most of client software display attached images as it is. It difficult for classifiers to detect image spam. There are two categorization as given below [18]

1. Emails messages are spam but it contains nonspam images
2. Email is spam and image itself contain spam text

OCR (optical character reorganization) techniques, signature base techniques are available for image filtering promising better techniques to improve image spam classification.

In our approach, we have used only header fields for classifying emails.

3. CLASSIFICATION TECHNIQUE FOR SPAM EMAILS CLASSIFICATION

Static method

Static method covers blacklist (predefined address list), and relies on rule based methods. This is the list of addresses from which emails are rejected, which gradually keeps building. These methods require continuous monitoring of incoming traffic and frequent updation of rules, which is one of its drawbacks.

Dynamic method

Dynamic method represents use of machine leaning algorithm such as Naïve Bayes, SVM, Decision tree, Random forest, k-nearest neighbor techniques.

Machine learning techniques have following benefits as compared to rule base learning techniques,

1. Accuracy.
2. Speed: Automatic classification makes machine learning faster and easy.
3. No need of frequent updation of rules
4. Lot of flexibility.

Bayesian network Classifier

Bayesian network is directed acyclic graphical model which allow representing probability distribution over set of

random variables [19]. Variable is represented as node. Conditional dependencies of the variables are represented by edges of the graph. Conditionally independent nodes are not connected to each other.

Consider finite set S={X1, X2,... Xn} set of discrete random variables. Each variable Xi, takes set of values.

let P is the joint probability distribution over variables in S and let X,Y,Z is subset of 'S', which takes values x,y,z respectively.

so, X and Y are conditionally independent given Z if for all x∈ val(X), y∈ val(Y), z ∈val(Z),

P(x|z, y)=P(x|z) whenever P(y,z)>0

Decision tree

Decision Tree (DT) is technique which is commonly used. The decision trees algorithm constructs a tree with graphical representation. J48 is an algorithm that builds decision trees from a set of training data using the concept of Information Entropy.

Decision tree is constructed by using training set with predefined set of classes. The features of known samples are applied to find properties of unknown samples. Decision tree provides good accuracy for large amount of data. Sometime it may represent complex structure.

Random forest

It is algorithm developed by Leo Breimanand Adele. It uses randomized node optimization. Ensembles are a divide-and-conquer approach used to improve performance. It takes N nodes at a time randomly to crate subset of data. Predictor variables are selected from data and the variables that give best split are used for that node. At next node next set of predictor variables are selected. These steps are repeated till all nodes a visit completes.[20]

K-nearest neighbor

K-nearest neighbor (KNN) finds out unknown data point based on known nearest predefined class. KNN calculates mean value of K nearest neighbor. Whenever new instance to classify, it calculates its k nearest neighbor from training data. Distance is calculated using Euclidean Distance. For example, an instance xq to be classified as, Let x1, x2….xk denote the k instances from training samples that are nearest to xq. Find the K nearest neighbors based on the Euclidean distance. It returns the class that represents the maximum of the k instances.

Bagging

Bagging is machine learning algorithm design to improve stability and performance. It helps to reduce overfitting and variance.

If 'Sm' is set of training emails with size 'n', than bagging generate new training dataset 'Smi' of size 'n' by sampling Sm uniformly. Small change in training data can significantly change the model [21]

4. DATA COLLECTION

Collection of Emails to be used as input to the processes was the first step. They were collected by programs written as part of current work using *JAVA mail API*. Emails in personal inbox were used as a sample data for extraction into a form which could be submitted as input to the various algorithms chosen. These data are used for training and testing purpose in email classification. We extracted a total of 849 Emails from the period June 2015 to March 2016. Out of that 592 are ham emails and 257 are spam. In order to obtain a training corpus for supervised learning algorithms, emails were classified as spam and nonspam.

In our approach we have used in all 11 header fields. Purpose was to find out whether using only header fields can lead to classification into ham or Spam. The features selected were

F(s)={CC, Date, Delivered To, DKIM Signature, From, Message ID, Reply To, Return Path, Received, Subject, To}

Thus our method had:

Input: E = email message

Output: C ∈ {spam, not-spam}

Objective: obtain a predictor f where, 'f' is called a classifier, and C is called as category.

Table -1: Meaning of email header fields

CC	Address of receiver/s intends to get messages
Date	orig-date ("Date" ":" date-time) when message sent
Delivered To	message received by machine
DKIM Signature	DomainKeys Identified Mail checks email contents and a message gets associated to a domain name
From	Senders address
Message ID	Unique identification number of message. It is composed of name of server that assign id and unique string
Reply To	Specifies address where sender wants replies to go
Return Path	If the message is rejected, it will be sent back to the email address listed here
Received	Contains information about receipt of the current message by a mail transfer agent on the transfer path.
Subject	Specifies topic description
To	Address of receiver

Collected emails were also labeled manually as spam and ham, for further use. To label message as spam message following rules are considered,

1. If 'from' field is empty or contain multiple numbers in the address field
2. 'Message id' may contain multiple '@' symbol or sender id domain and message id domain may different.
3. 'Subject' contains invalid Text or some common term such as "lottery", "you won"
4. 'Return-path' contains some invalid address or username. For example, yyycccxx@gmail.com
5. In 'Received' field 'with' tag missing or sender addresses pretending to be root user.

5. RESULT AND DISCUSSION

Classifier provided in Weka were used for classification, this was possible since all emails were extracted into a proper flat data base using the codes written for the purpose, as mentioned earlier. N Fold cross validation technique is used. In experiment we have considered N=10. It will divide entire dataset into 10 parts, for each iteration 9 portions of datasets is used as training and 1 part is used for testing. This is repeated 10 times with different portion of dataset as training and testing. Five types of classifiers were tested, Bayes Net, K-nearest neighbor, Decision tree and Bagging.

Classifiers	Time	% Emails Correctly Classified	% Emails Incorrectly Classified	Tp Rate	Fp Rate	Precision	Recall
BAYES NET	0.05 seconds	97.87% (830)	2.12% (19)	0.979	0.003	0.997	0.979
K-Nearest Neighbor (Lazy.IBK)	0.02 seconds	99.29 (842)	0.70% (7)	0.993	0.013	0.992	0.993
Decision Tree (J48)	0.48 seconds	99.64 (845)	0.36% (4)	0.996	0.008	0.994	0.996
Random Forest	0.09 seconds	73.58% (624)	26.42% (225)	0.736	0.601	0.806	0.736
Bagging	0.3 seconds	69.69% (591)	30.31% (258)	0.69	0.68	0.78	0.69

Table2. Comparative output of classifiers

As per output shows in table1, decision tree and K-nearest neighbor algorithm outperforms with header features. Bagging shows poor performance among all the classifiers. Decision tree took more time to build model. Bayes Net achieves highest precision. Decision tree also shows highest true positive Classifier performance is represented using Accuracy, True Positive (TP), False Positive (FP), Precision, and Recall.

True positive (TP) Rate: Actual class and predicted class are same.

False positive (FP) Rate: Predicted to be in class but does not belong to that class.

Accuracy (correctly classified): Total number of test records correctly classified by model.

Incorrectly Classified: Number of records incorrectly classified by model

Precision can be seen as a measure of exactness or quality,

Recall is a measure of completeness or quantity.

$$\text{Recall} = \frac{\text{Number of emails correctly classified as positive}}{\text{Number of positive email records}}$$

$$\text{Precision} = \frac{\text{Number of correct positive predictions}}{\text{No of positive predictions}}$$

$$\text{Accuracy} = \frac{\text{Correctly Classified Emails}}{\text{Total Emails}}$$

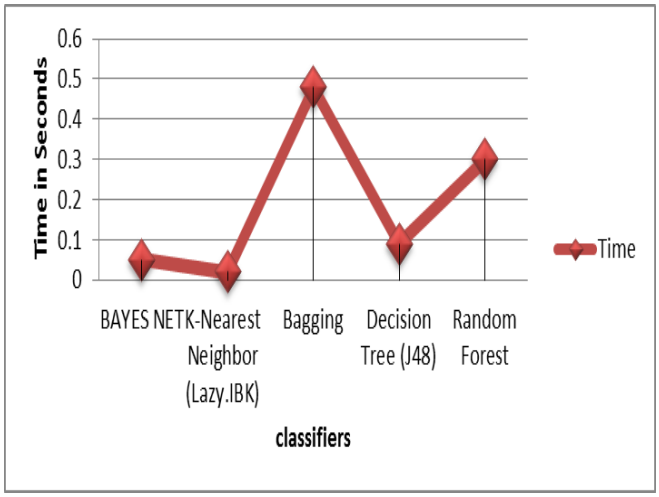

Chart -1: Time taken by classifiers

5.1 Comparison of our result with existing literature result

Author/paper	Classifier	Dataset	Accuracy(%)
Trivedi S and Dey S (2013) [25]	Probabilistic classifiers with Boosting	Enron Email dataset	92.9
Metsis et al. (2006) [23]	Five types of Naïve Bayes comparison	Enron data set with different compositions	90.5 to 96.6
W.A. Awad, and S.M. Elseuofi (2011) [24]	Bayesian, SVM and various ML	Spam Assassin	97.42-99.46
Youn, S. a. (2006)[22]	using J48	***	95.80%

6. CONCLUSIONS

In this paper, Decision tree, Bayes network, K-Nearest Neighbor, Random Forest and Bagging algorithms are used to test spam classification using email header fields. Result shows that decision tree (J48) is very simple and performs better than all classifiers. K-nearest neighbor also performs good but bagging and random forest does not show promising result.

We have proved that using email header emails can be classified as spam and non-spam to certain extent. Result is compared with existing result of content base classification .

REFERENCES

[1] Gordon Cormack and Thomas Lynam.(2005). Spam corpus creation for TREC. In Proceedings of Second Conference on Email and Anti-Spam, CEAS'2005, 2005.
[2] http://www.monkeys.com/spam-de_ned/
[3] https://usa.kaspersky.com/internet-security-center/threats/spam-statistics-report-q1-2014#.VvKvw-J97Dc
[4] Justin M. Rao and David H. Reiley, The Economics of Spam, Journal of Economic Perspectives—Volume 26, Number 3—Summer 2012—Pages 87–110
[5] http://www.spamlaws.com/spam-filters.html
[6] http://www.statista.com/statistics/420391/spam-email-traffic-share/
[7] Patidar, V., Singh, A., & Singh, A. A Novel Technique of Email Classification for Spam Detection.
[8] Enrico Giacoletto and Karl Aberer,"Automatic Expansion of Manual Email Classifications Based on Text Analysis",LNCS 2888, pp. 785–802, 2003.© Springer-Verlag 2003
[9] Seongwook Youn, Dennis McLeod, "Spam Email Classification using an Adaptive Ontology" JOURNAL OF SOFTWARE, VOL. 2, NO. 3, SEPTEMBER 2007
[10] MD. Rafiqul Islam and Morshed U. Chowdhury, "Spam Filtering Using ML Algorithms", IADIS International Conference on WWW/Internet 2005, pp. 419-426.
[11] Basavaraju, M., & Prabhakar, D. R., "A novel method of spam mail detection using text based clustering approach," International Journal of Computer Applications, vol. 5, no. 4, pp. 15–25, August 2010
[12] Provost, Jefferson. "Naıve-bayes vs. rule-learning in classification of email."University of Texas at Austin (1999).
[13] Liu Pei-yu, Zhang Li-wei and Zhu Zhen-fang, "Research on Email Filtering Based on Improved Bayesian", Journal of Computers, v4, 2009, pp. 271-275.
[14] Le Zhang, Jingbo Zhu, and Tianshun Yao. An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3(4):243{269, 2004. ISSN 1530-0226.
[15] J.-J. Sheu, "An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization" International Journal of Network Security, vol.9, pp. 34-43, July 2009.
[16] Wang, Min-Feng, et al. "Enterprise email classification based on social network features." Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011.
[17] Ke, Liyiming, Bo Li, and Yevgeniy Vorobeychik. "Behavioral Experiments in Email Filter Evasion." (2016).
[18] Sheena, Preeti Anand, filtering image spam-a survey, International Journal of Computer Applications (0975 – 8887) Volume 79 – No5, October 2013
[19] Friedman, N., Geiger, D., & Goldsmith, M. (1997). Bayesian network classifiers. Machine learning, 29(2-3), 131-163.
[20] https://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics/
[21] https://en.wikipedia.org/wiki/Random_forest

[22] Youn, S. a. "A Comparative Study for Email Classification. "JOURNAL OF SOFTWARE , 2 (3), 1-13 (2006)

[23] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes – Which Naive Bayes?" in Third Conference on Email and Anti-Spam (CEAS 2006), USA, 2006

[24] W.A. Awad and S.M. Elseuofi, "Machine Learning Methods For Spam E-Mail Classification", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011

[25] S. K. Trivedi and S. Dey, "Effect of feature selection methods on machine learning classifiers for detecting email spams," in Proc. The ACM-SIGAPP Research in Adaptive and Convergent Systems (ACM RACS 2013), pp. 35-40, 2013