International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Mail Spam Detection Using Clustering & Random Forest Algorithm

Neel Varun[1], Pratap Singh[2] , Keshav Agrawal[3]
Department Of Computer Science & Engineering
Jaypee University Anoopshahr
neelgupta003@gmail.com[1], singhpratap.engr@gmail.com[2]

communication but unsolicited emails hamper such communications. Spam emails are considered as a serious violation of privacy , Spam filtering has become a very important issue throughout the last years as unsolicited bulk e-mail imposes large problems in terms of both the amount of time spent on and the resources needed to automatically filter those messages .The present research emphasises to build a spam classification model with/without the use of Clustering messages that allows for efficient labeling of a representative sample of messages for learning a spam detection model using a Random Forest for classification . This paper described classification of emails by Random Forests (RF) Algorithm. RF is ensemble learning technique. The Random forest is a meta-learner which consists of many individual trees. Each tree votes on an overall classification for the given set of data and the random forest algorithm chooses the individual classification with the most votes. If identified category is 0 then e-mail is marked as non-spam email otherwise if identified category is 1 then e-mail is marked as spam e-mail . Through this study, the aim is to distinguish between ham emails and spam emails by making an efficient and sensitive classification model that gives good accuracy with low false positive rate.

1. INTRODUCTION

Internet started to gain popularity in the early 90''s, it was recognized as an good advertising tool. With out any cost, a person can use the Internet to send an email message to many people. When this message contains an unwanted Advertisement, it is commonly known as spam email.Undesired email is a nuisance for its recipients; however, it also often presents a security threat. For example, it may contain a link to a fake website wanted to capture the user's login or personal detail(identity theft, phishing), or a link to a website that contain malicious software (malware) that can damage user's computer. These malware can be used to capture user information, send spam, host malware or conduct service attacks as part of a "bot" net. While prevention of spam transmission would be ideal, detection allows users and email providers to address the problem today. Spam detection programs use some keywords like guaranteed, free etc and block any email with those words in them. But this has the disadvantage of sometimes blocking even important mails from your contacts and preventing those senders from sending mails to your address again. The way out is to use add-on spam filters which allow you to control the content that should be allowed into your inbox. This will save you a lot of time and energy as you no longer will have go through each and every email before identifying it as spam and eliminating it. Given below is the statistics of global spam volume as percentage of total e-mail traffic as of September 2018, sorted by month. As of the most recently reported period, spam messages accounted for 53.5 percent of e-mail traffic worldwide. In the second quarter of 2018, China accounted for the majority of unsolicited spam e-mails with 14.36 percent of global spam volume. The most common types of spam e-mail were healthcare and dating spam [6].

Table1 Global spam volume as percentage of total e-mail traffic from January 2014 to September 2018, by month

MONTH	PERSENTAGE OF EMAIL SPAM TRAFFIC
Jan 14	65.7%
Feb 14	69.9%
Mar 14	63.5%
Apr 14	71.1%
May 14	69.9%
Jun 14	64.8%
Jul 14	67%
Aug 14	67.2%

2. RELATED WORK

Zhan Chuan[1] et al proposed An Improved Bayesian with Application to Anti-Spam Email in which they presents a new improved Bayesian-based anti-spam e-mail filter. They adopt a way of attribute selection based on word entropy, use vector weights which are represented by word frequency, and deduce its corresponding formula. It is proved that their filter improves total performances

apparently. Denil Vira[2] et al present An Approach to Email Classification Using Bayesian Theorem. They propose an algorithm for email classification based on Bayesian theorem. The purpose is to automatically classify mails into predefined categories. The algorithm assigns an incoming mail to its appropriate category by checking its textual contents. The experimental results depict that the proposed algorithm is reasonable and effective method for email classification. Vikas P. Deshpande[3] et al proposed An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques in which efficient anti-spam filter that would block all spam, without blocking any legitimate messages is a growing need. To address this problem, they examine the effectiveness of statistically-based approaches Naïve Bayesian anti-spam filters, as it is content-based and self-learning (adaptive) in nature. Additionally, they designed a derivative filter based on relative numbers of tokens. They train the filters using a large corpus of legitimate messages and spam and also test the filter using new incoming personal messages. Mehran Sahami[4] et al examine methods for the automated construction of filters to eliminate such unwanted messages from a user's mail stream. By casting this problem in a decision theoretic framework, they are able to make use of probabilistic learning methods in conjunction with a notion of differential misclassification cost to produce filters. In order to build probabilistic classifiers to detect junk Email, they employ the formalism of Bayesian networks. Denil Vira[5] et al present An Approach to Email Classification Using Bayesian Theorem. They propose an algorithm for email classification based on Bayesian theorem. Georgios Paliouras & Vangelis Karkaletsis present Learning to Filter Spam E-Mail A Comparison of a Naïve Bayesian and a Memory-Based Approach in which they investigate the performance of two machine learning algorithms in the context of anti-spam Filtering. They investigate thoroughly the performance of the Naïve Bayesian filter on a publicly available corpus, contributing towards standard benchmarks. At the same time, we compare the performance of the Naive Bayesian filter to an alternative memory based learning approach, after introducing suitable cost-sensitive evaluation measures. Both methods achieve very accurate spam filtering, outperforming clearly the keyword-based filter of a widely used e-mail reader.[8].

3. MATERIAL AND METHODS

Majority of the email spam filtering methods uses text categorization approaches. Consequently, spam filters perform poorly and cannot efficiently prevent spam mails form perform poorly and cannot efficiently prevent spam mails from getting to the inbox of the users. This work employes , rules using Random Forests (RFs) algorithm to extract important features from emails, and classify the emails into either ham or spam .The data used for this project was taken from the Spam Assassin public corpus website. It consists of two data sets: train and test. Each dataset contains a randomly selected collection of emails in plain text format, which have been labelled as HAM or SPAM. The training data is used to build model for the classifying emails into HAM and SPAM. The test data is used to check the accuracy of the model built with the training data. The training data set contains 1000 emails with 500 ham and 500 spam emails. The test data contains 200 emails with 139 ham and 61 spam emails.

4. CLUSTERING

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. In this project, clustering used to select an initial set of email messages to be labeled as training examples.In this project PAM clustering algorithm is used PAM stands for “partition around medoids”. The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters. Objects that are tentatively defined as medoids are placed into a set S of selected objects. If O is the set of objects that the set U = O − S is the set of unselected objects. The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object.

The algorithm has two phases:

(i) In the first phase, BUILD, a collection of k objects are selected for an initial set S.

(ii) In the second phase, SWAP, one tries to improve the quality of the clustering by exchanging selected objects with unselected objects.

Equivalently, we can minimize the sum of the dissimilarities between object and their closest selected object. The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object. Equivalently, we can minimize the sum of the dissimilarities between object and their closest selected object.

5. RANDOM FOREST

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results. The random forest starts with a standard machine learning technique called a “decision tree” which, in ensemble terms, corresponds to our weak learner. In a decision tree, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. Random forest runtimes are quite fast, and they are able to deal with unbalanced and missing data. Random Forest weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may

over-fit data sets that are particularly noisy. Of course, the best test of any algorithm is how well it works upon your own data set. Random Forests grows many classification trees. Each tree is grown as

i. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
ii. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
iii. Each tree is grown to the largest extent possible. There is no pruning.Table 2 shows a comparison of cross validation performance using 10 cluster prototypes for training.

The performance measure is Area Under the receiver operating characteristic Curve (AUC).

Table 2. Comparison of Cross Validation Performance

Method	ACU
Random Forest(proposed)	95%
Naive Bayes[4]	66.7%
SVM[4]	66.7%
KNN[4]	66.7%

6. CONCLUSION

Various existing email spam detecting technique are not much effective to some of the spams been sent by spammers. This is because spammers kept on developing new complex techniques for evading detection by spam detector. With continuous usage of new technique by spammers, email spam filtering has become a hot research area for researchers .study we proposed Random Forests algorithm for effective and efficient email spam filtering. And evaluated the performance of RFs algorithm.We conclude that RFs is a promising algorithm that can be adopted either at mail server or at mail client side to further decrease the volume of spam messages in email users inbox.

REFERENCES

[1] Zhan Chuan, LU Xian-liang, ZHOU Xu, HOU Meng-shu,"An Improved Bayesian with Application to Anti-Spam Email", Journal of Electronic Science and Technology of China, March 2005, Volume 3, Issue 1.
[2] Denil Vira, Pradeep Raja & Shidharth Gada,"An Approach to Email Classification Using Bayesian Theorem", Global Journal of Computer Science and Technology Software & Data Engineering Year 2012,Volume 12 ,Issue 13 Version 1.0
[3] Vikas P. Deshpande, Robert F. Erbacher, "An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques", Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy, West Point, NY 20-22 June 2007.
[4] Mehran Sahami ,Susan Dumaisy, " A Bayesian Approach to Filtering Junk E-Mail ",GatesBuilding 1A Computer Science Department Microsoft Research Stanford University Redmond, WA 98052-6399,Stanford, CA.
[5] Denil Vira, Pradeep Raja & Shidharth Gada,"An Approach to Email Classification Using Bayesian Theorem", Global Journal of Computer Science and Technology Software & Data Engineering Year 2012,Volume 12 ,Issue 13 Version 1.0
[6] https://www.statista.com/statistics/420391/spam-email-traffic-share/
[7] Fawcett, T. "An Introduction to ROC Analysis", Pattern Recognition Letters, Vol 27, Iss 8, Jun 2006, pp. 861-874.