Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 9; April-June, 2015 pp. 64-65

http://www.krishisanskriti.org/acsit.html

Performance Comparison of Machine Learning based Spam Filtering

Nadia Nisar

M.Tech Scholar, Department of Computer Science & Engineering Al-Falah School of Engineering and Technology Dhauj, Faridabad, Haryana
E-mail: nadia.nisar786@gmail.com

Abstract—*Machine Learning approach has been widely used in spam filtering and is regarded as an efficient approach towards filtering the unsolicited e-mails. However, the machine learning approach is considered viable only to the e-mail body contents excluding the e-mail header section. Little effort has been put towards the comparative analysis of variations possible in spam filtering approach based on machine learning. In this paper we analyse the performance of machine learning based spam filtering technique applied on both e-mail headers and e-mail body. We compare the proposed technique with the classical filtering approach which utilizes only the e-mail body for spam identification. Comparison has highlighted the significant increase in true positives and reduction in false positives.*

Keywords*: Spam, Naive Bayes, Weka, Classifier, Dataset*

1. INTRODUCTION

Spamming may be defined as flooding the internet with hundreds and millions of unsolicited bulk messages. People spend a lot of time and effort in order to get away with the unsolicited emails received as a substantial part of the received e-mails in the user inbox. The frequency and the magnitude of the received spam messages can sometimes be of several orders of magnitude in comparison to the solicited e-mail messages. Such a transmission leads not only to the wastage of bandwidth but is a serious security concern as well, hence a huge burden on the organisation's profitability as well as credibility. Further, spams can cause various security issues like phishing and spoofing attacks which can lead to money laundering and other monetary losses. The current era of technological revolution is highly dependent on the effectiveness of the underlying communicating media, thereby highlighting the importance of detecting and filtering non-legitimate emails. There have been a number of contemporary approaches to counter it, but none of them is fool proof. Spammers can use malware combined with the power of botnets to launch large scale spamming missions causing major traffic increase and leading to enormous economical loss. With every new technique evolved in the market to filter spam emails, spammers in turn develop pioneering approaches to bypass them . This signifies the importance of performance study of various machine learning based spam filtering techniques in order to compare their efficiencies.

2. MOTIVATION

It has been analyzed that the spam filtering done on the principles of Message body semantics is time consuming and rigorous process. Modern day spam filtering techniques also include rules based on information contained in the email headers. There are various fields in e-mail headers that could be utilized to classify spams. For example, header information can reveal if the email is from a recognizable domain that is associated to the actual sender name. Finding out the IP reputation using Return Path E-mail Header is another example to utilize header information for spam classification. Although, extensive work has been done on machine learning approach applied to e-mail contents (e-mail body), little research has been done on the application of classifiers like naive-bayes on e-mail headers. In contemporary spam filters Header based filtering is used in conjunction with e-mail body semantics based techniques. However, novel machine learning principles applied on e-mail headers could simply the filtering process and avoid rigorous and time consuming techniques applied on e-mail body. A comparative performance study of machine learning based spam filtering used on e-mail headers and e-mail body contents will help analyse their efficiencies.

3. RECENT ADVANCES

Both the sections of the e-mail message could hold valuable information for a spam filter to classify a mail as either a spam or a non-spam. An extensive research based on whitelisting/blacklisting, Bayesian probability analysis, mail body keyword tokenization and checking etc have already being conducted by many researchers. One of the common methods for spam detection is Bayesian spam filtering (Thomas Bayes) which is a statistical technique for e-mail filtering. In this technique a naïve Bayes classifier is used to identify spam e-mail. Bayesian classifier work by correlating the use of tokens with spam and non-spam e-mails and then using Bayesian inference to calculate a probability that an e mail is spam or not.

4. PROPOSED METHODOLOGY

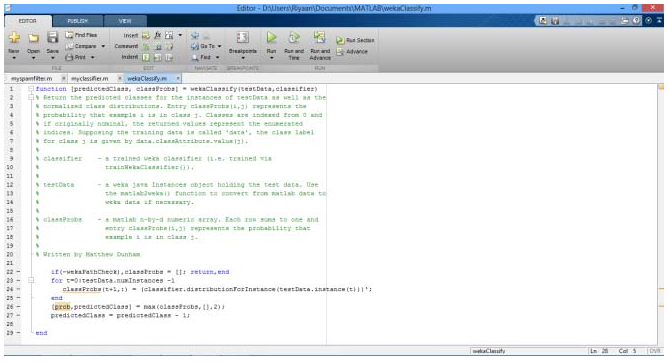
Machine learning techniques used for spam filtering use email messages body to predict whether the mail is spam or not. Open source spam filtering datasets are widely available for the use as a testing and training datasets for machine learning approach. It should be noted that such datasets available are based upon the features/attributes extracted from the e-mail body and not from the e-mail header. Spambase is one such dataset offered by UCI –machine learning repository. In order to compare the header based machine learning technique with contemporary machine learning approach we need to make a header based dataset comprised of attributes extracted from e-mail headers. The attributes of the header can be converted and used to create a training and subsequent testing dataset.

Implementation

Our interface in matlab is myspamfilter.m program. In this program we first extract the header value pairs in the header of our mail by making ':' as delimit. Splitstring() function is used to split the header attributes. Then we tokenize the body of the mail according to the words described in spambase.arff file. We count the word frequency using arff_body_parser( ) function. This program gives us an array of data or feature set which is stored in our testspambase.arff file for further classification.

After header value separation of header attributes and tokenization of words in body of the mail, two files that are test.arff and training.arff are made arguments to the function myclassifier(a_test,a_training) which classifies the test data using naïve bayes algorithm and predicts that whether the mail is spam or not. This classification is done through weka using wekaClassify() function. As described weka is an open source classification application whose input is an arff file i.e attribute relation file format. Weka returns various values like mean, standard deviation, error rate etc and also predicts whether a mail is spam or not.

wekaClasiffy(testData,classifier) is the actual function which gets the predicted probability and predicted class from weka by giving it test.arff and training object. The values are calculated in weka and are visualized in matlab.



A custom made training dataset on the basis of feature set identified from the e-mail header fields was created using 500 e-mails. The training dataset was merged with simplified variant of spambase dataset.

MATLAB platform was used to test the spam filtering methodology in a passive manner. Further, Matlab was used to calculate and compare the confusion matrix in both the scenarios. The matrix is based on set of test dataset (ARFF File) made up of 50 mails (1:1 ratio). The Matrix suggests no evidence of performance degradation with simplified spambase dataset. However, significant amount of time was saved to suggest that header based machine learning can reduce the time required to classify the mails. The subsequent integration with a Mail User Agent could be done to test the performance in the active manner.

Table 1: When only spambase dataset was used

	Actual	
Predication	Spam	Non-Spam
Spam	19	4
Non-Spam	6	21

Table 2: When simplified spambase was used with custom header based dataset

	Actual	
Predicition	Spam	Non-Spam
Spam	18	3
Non-Spam	7	22

5. CONCLUSION

When the Naïve-Bayes classification technique is applied using machine learning based dataset encompassing certain header field based attributes, the rigorous and time consuming filtering done on e-mail body can be simplified without affecting the spam filter performance. The conclusion was arrived after removing certain attributes and simplifying others from spambase dataset and incorporating the custom developed dataset.

REFERENCES

[1] Shukor Bin Abd Razak, Ahmad Fahrulrazie Bin Mohamad Identification of Spam Email Based on Information from Email Header, 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA)

[2] Understanding an email header http://kb.mediatemple.net/

[3] Spamming Botnets: Signatures and Characteristics Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten+, Ivan Osipkov+ Microsoft Research, Silicon Valley

[4] Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304