ISSN: 0974-6471, Vol. 10, No. (3) 2017, Pg. 698-702

Oriental Journal of Computer Science and Technology

Journal Website: www.computerscijournal.org

Performance Analysis, Comparative Survey of Various Classification Techniques in Spam Mail Filtering

DR. PRITI[1] and UMA[2]

[1]Assistant Professor at D.C.S.A. ,M.D.U. ,Rohtak , India.
[2]Research Scholar at D.C.S.A. ,M.D.U., Rohtak , India.

Abstract

One of the most common methods of communication involves the use of e-mail for personal messages or for business purposes. One of the major concerns of using the e-mails is the problem of e-mail spam. The worst part of the spam e-mails is that, these are invading the users without their consent and bombarding of these spam mails fills up the whole e-mail space of the user along with that, the issue of the wasting the network capacity and time consumption in checking and deleting the spam mails makes it even more concerning issue. With the increasing demand of removing the e-mail spams the area has become magnetic to the researchers. This paper intends to present the performance comparison analysis of various pre-existing classification technique. This paper discusses about spam mails in section (I), In section (II) various feature selection methods are discussed , In section (III) classification techniques concept in spam filtering has been elaborated, In section (IV) existing algorithms for classification are discussed and are compared. In section (V) concludes the paper giving brief summary of the work.

Article History

Received: 09 June 2017
Accepted: 28 June 2017

Keywords

Classification,
E-mail Threats,
Spam Filtering,
Efficiency ,
Feature selection.

Introduction

With the most preferred communication method e-mails have become part of day to day life. Spams which are also called unwanted, junk ,unsolicited mail is one of the major problem in using the e-mails. There are basically two things that are confused with each other that are one is Paper junk mail and spam mail, these two are usually confused with each other. Let's clear this concept that in the Paper Junk Mail Junk mailers pay for distribution of the material while in case of E-mail spamming the recipient has to pay in terms of bandwidth, disk space, server resources as well as lost productivity. The issue of e-mail spamming can become a headache if not managed properly[1]. There are many issues that arise from the bombardment of the

CONTACT UMA ✉ umabhardwaj90@gmail.com Research Scholar at D.C.S.A. ,M.D.U., Rohtak , India.

To link to this article: http://dx.doi.org/10.13005/ojcst/10.03.21

spam e-mails like filling up of the user's mailboxes, flooding important e-mails, wastage of memory along with bandwidth and time.

What is Spam?

When the question comes regarding what Spam actually is it can be defined as the unwanted and unsolicited e-mails that come from strangers and are broad casted to multiple number of email-ids[1]. Spam is the engulfing of the internet in which many copies of the same message sent to people who would not choose to receive it. Mostly Spam mails are used for doubtful products, get rich quick schemes.

Spam Filtering

Spam filtering is a process that is used to detect unsolicited and unwanted e-mails and prevent those messages from getting to a user's inbox. There are two levels at which the Spam filtering in the e-mails can operate that will involve a user level or an enterprise level. Individual Users refers to the single specific person that is working at home and who has been receiving and sending the e-mails via ISP, these users if wish to identify and filter the spam mails simply install the spam filtering system. In the Enterprise level spam filtering mails are filtered during entering time in the internal network of an Enterprise. In the Enterprise level spam filtering spam filtering software is installed on the main mail server and it is meant to interact with the mail transfer agent (MTA) that classifies the message at the moment they are received[1]. Most by far of current spam sifting frameworks use principle based scoring systems. An arrangement of tenets is connected to a message and a score gathers in light of the guidelines that are valid for the message. Frameworks commonly incorporate several guidelines and these standards should be redesigned frequently as spammers modify substance and conduct to maintain a strategic distance from the channels. The engineering of spam separating is shown in Fig.2. Initially the model will collect the client messages which can be spam mail and non-spam mail. Then the underlying change procedure will start. The model states starting change, the user identification, highlight extraction, e-mail information order, analyzer area.

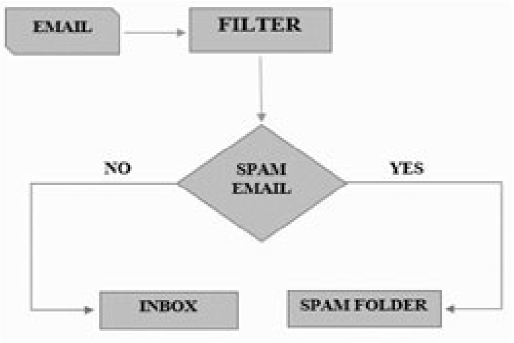

Fig. 1: Flowchart of Spam mail Filter

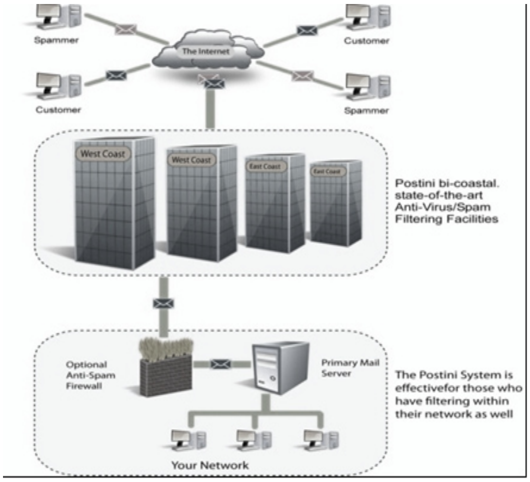
Spammer
Customer
The Internet
Customer
Spammer
West Coast
West Coast
East Coast
East Coast
Postini bi-coastal. state-of-the-art Anti-Virus/Spam Filtering Facilities
Optional Anti-Spam Firewall
Primary Mail Server
The Postini System is effectivefor those who have filtering within their network as well
Your Network

Fig. 2: The Process of Spam Mail Filtering.

Feature Selection Methods

Feature selection is also termed as attribute selection. It automatically select relevant attributes from the data which are used in predictive model construction. It is a technique which reduce the number of inputs for processing. It removes the extra attributes which reduce the accuracy of the model. As mentioned in[9] , Information Gain , Gini Index , Term Frequency Inverse Document Frequency are some most popular feature selection techniques.

Information Gain (IG)

It is a method used by Decision tree for attribute selection .The attribute which have highest value of information gain used as splitting attribute. The large value of information gain for a attribute makes it more significant.

Gini Index (GI)

It uses the binary split for splitting of a attribute. In case of discrete value attribute minimumgini index value is selected as splitting criteria. In case of continuous attributes every split point will be considered.

Gain Ratio (GR)

It uses the normalization on the information gain technique. The attribute with the maximum gain ratio is used as splitting attribute.

Fuzzy Adaptive Particle Swarm Optimization (FAPSO)

It works on three levels , core feature subset selection , Feature subset selection , and spam filtering. It finds the relevant feature from data set.

Term Frequency Inverse Document Frequency

It is a mathematical technique. It finds the frequency of a word in a document. It calculates the importance of a word in a document. The words which are frequently used have high value of TF-IDF.

Classifiers In Spam Mail Filtering

There are many types of classifiers that are meant for the purpose of classifying the e-mails as spam or not and these are basically classified into two categories mainly those being: Content based classifiers and Non-content based classifiers.

Content Based Classifiers

These classifiers are also famous by the name of hand crafted spam classifiers and these are the types in which the spams are categorized on the basis of the content it holds or information it stores. It checks for text in body of the E-mail, then URL. It also considers the mail header like subject for classification of text. It performs text classification task by employing preprocessing on text in terms of HTML tags removal, Tokenizing and Word frequency calculation for determining word probability to find out whether a given mail is spam or not.

Non-content based classifiers

In this type of the classifier the automated filter is installed and in this the classification depends upon the human recipient. In this the classification occurs from the judgment of the sender's name, address etc.

Types Of Classification Algorithms

There are many algorithms that are designed for the purpose of e-mail classification and some of them are discussed below:

Naive Bayes Algorithm

It is one of the famous machine learning algorithm working on the principle of Bayes theorem. Bayes theorem calculate the posterior probability. It is the technique that is widely used for the purpose of e-mail classifications for spam and non-spam. It is defined as:

$$P\ (H/K) = P\ (K/H)\ P\ (H)\ /\ (P\ (K). \qquad ...(1)$$

Where,

P (H/K) is the posterior probability of class (H) for given predicator (K).
P (K/H) is the likelihood which is probability of predicator for given class.
P (H) is the prior probability of class.
P (K) is the prior probability of the predicator.

Some common words are used in both spam and non-spam mails. It is not like that filters know the words previously, but there has to be a training process built up for them and after that these word probabilities are utilized for the purpose of e-mail classifications. In this case, each word or the most interesting words contribute to the e-mail spamming. And there is a threshold that has been set for determining the spam and if the probability is increased above that threshold, then the e-mail is considered as the spam.[2,3,4]

Support Vector Machine Algorithm

SVM is a supervised machine learning technique which is used for both classification and regression. In this we plot each data item as a point in n-dimensional space where :-
n= number of features.
Then it performs classification by finding the hyper - plane that differentiate the two different classes.[5,6]

k-Nearest- Neighbor Algorithm

The k-Nearest Neighbor (kNN for short) is a non-parametric instance based learning technique or lazy learning. It is used for make decision based on complete training data set. The input consists the k closest data items in the feature space. The output is a class membership function. An object is classified by majority vote of its neighbors. The object will be assigned to the class which is most common among k nearest Neighbors.[8]

Decision Tree Induction Algorithm

Decision tree consist the root node, branches and leaf nodes. In this the tree is created in a top-down, recursive and divide and conquer way. It works like a greedy technique. The internal node defines the condition on the attribute, each branch defines the output of the condition and each leaf node defines the class label.[9]

Rule Based Classification Algorithm

In the algorithm classifier is represented as a set of IF-THEN rules. IF-THEN rule is of the form IF condition THEN conclusion. The "IF" part is called as rule antecedent. The "THEN" part is called as rule consequent. The condition performs the test on one or more attributes. The class prediction are specified by rule consequent.

Back propagation Algorithm

It is a neural network learning algorithm. It trains the feed forward multi layer neural network for given data samples. When each entry of the sample data item is presented to the network, the network checks the output response to the sample data item. The output response is then compared with known and desired output and error value is find out. Based on error value network weights are adjusted. The weights are adjusted by finding mean square error of output response with input sample.[7] principles of each classification technique is highlighted with their findings and limitations.

Table 1: Theoratical Findings of Classification Techniques

Sr. no	Algorithm	Classification	Findings Principle	Limitations
1.	Naive Bayes Algorithm	Works on Bayes Theorem.	It has high accuracy and speed when used for large data sets.	Assumption is made that events occurring are mutually exclusive.
2.	Support Vector Machine Algorithm	Non- Linear Mapping.	Highly Efficient and accurate classifier. Less prone to over fitting.	Complex algorithm difficult to understand. Training time is more.
3.	k-Nearest-Neighbor Algorithm	Learning by analogy and distance based comparison.	Less work on training data sets but more work on classification.	Computationally expensive. Require efficient storage techniques.
4.	Decision Tress Induction Algorithm	Top down, recursive, divide and conquer based.	Can handle high dimensional data. It is simple and fast and have good accuracy.	Branches may contain outliers in the training data sets.
5.	Rule Based Classification	Based on IF-THEN rules.	Rules are efficient technique for the representation of knowledge. Rules are specified by using coverage and accuracy.	What if more than rules is fired and specify different classes. And if no rule is fired.
6.	Classification by Backpropagation	Based on neural network learning algorithm.	Can deal with noisy data and have capability to classify data sets for which they are not trained.	Require more training time. Suffers from Poor interpretability.

Conclusion

Efficiency of spam mail filtering is depending on classification algorithm used. In this paper, a number of existing algorithms for spam mail filtering are discussed, compared with each other and tabulated with their findings[12]. It helps to understand the wide variety of classification techniques in order to select one.

Acknowledgment

The paper has been composed with kind assistance, guidance and support of my department who have helped me in this work. We would like to thank all the people whose encouragement and support has made the fulfillment of this work conceivable.

References

1. Omar Saad, Ashraf Darwish and Ramadan Faraj: "Asurvey of machine learning techniques for Spam filtering", IJCSNS ,International Journal of Computer Science and Network Security, VOL.12 No.2, February 2012.
2. I. Androutsopoulos, J. Koutsias, "An evaluation of naïve Bayesian anti-spam filtering", 11thEuropean Conference on Machine Learning (ECML 2000),pp 9–17, 2000.
3. Androutsopoulos, G. Paliouras, "Learning to filter spam E-mail: A comparison of a naïve Bayesian and a memory based approach", 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp 1–13, 2000.
4. K. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering", 10th Conference of the European Chapter of the Association for Computational Linguistics, pp.307-314, 2003.
5. N. Cristianini, B. Schoelkopf, "Support vector machines and kernel methods, the new generation of learning machines". Artificial Intelligence Magazine, pp 31–41, 2002
6. S. Amari, S. Wu, "Improving support vector machine classifiers by modifying kernel functions". Neural Networks, pp 783–789, 1999.
7. C. Miller, "Neural Network-based Antispam Heuristics", Symantec Enterprise Security (2011), www.symantec.com Retrieved December 28, 2011
8. Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora , "Text and image based spam e-mail classification using KNN, Naïve Bayes and reverse DBSCAN Algorithm, " ICROIT 2014, India, Feb 6-8 2014.
9. Masurah Mohamad and Ali Selamat, "An evaluation on the efficiency of hybrid feature selection in spam e-mail classification," IEEE International Conference on Computer Communication, and Control Technology (14CT 2015), April. 2015.
10. Rushdi Shams and Robert E. Mercer, "Classification spam e-mails using text and readability features," IEEE 13th International Conference on Data Mining, pp. 657-666, 2013.
11. Megha Rathi and Vikas Pareek, "Spam E-mail Detection through Data Mining-A Comparative Performance Analysis," I.J. Modern Education and Computer Science, pp. 31-39, 2013.
12. Ms.D. Karthika Renuka, Dr.T. Hamsapriya, Mr.M. Raja Chakkaravarthi, Ms.P. Lakshmisurya, "Spam Classification based on Supervised Learning using Machine Learning Techniques," *IEEE*, pp.1-7, 2011
13. V. Vaithiyanathan , K. Rajeswari , Kapil Tajan , Rahul Pitale, "Comparison Of Different Classification Techniques Using Different Data sets" , IJAET , ISSN: 2231-1963 ,Vol. 6, Issue 2, pp. 764-768 , May 2013