A survey on spam email classification

Aawesh kumar borkar [1] , Ms. Anshul Singh[2]
M.Tech. Scholar, Department of Computer Science and Engg. RCET Bhial, Durg India
Assistant Professor, Department of Computer Science and Engg. RCET Bhilai, Durg India
[1] aawesh.desire28@gmail.com,
[2] anshul.31.singh@gmail.com

Abstract—**Nowadays the Email has become popular and inexpensive channel for communication. Due to increasing number of Email users there is resulted increase in number of spam email during past few years. In this paper we have discussed various spam classification techniques such as decision tree, Support vector machine, Bayesian classifier etc. We have also discussed their advantages and disadvantages. We have compared the various performance measures like accuracy, sensitivity and specificity of these methods.**

Keywords - **Spam, Classification, Neural network, Decision tree**.

I. INTRODUCTION

Spam [1] has become one of the biggest worldwide problems facing The Internet today. The Internet is becoming an integral part of our Everyday life and the e-mail has been a powerful tool for idea and Information exchange, as well as for users' commercial and social lives. Due to the increasing volume of spam, the users as well as internet service providers (ISPs) are facing a lot of problems. The cost to corporations in band width, delayed e-mail, and employee productivity has become a tremendous problem for anyone who provides e-mail services. However, despite the increasing development of anti-spam services and technologies it is amazing that, the number of spam messages continues to increase rapidly. E-mail classification techniques are able to control the problem in a variety of ways. Detection and protection of spam e-mails from the e-mail delivery system allows end-users to regain a useful means of communication.

Ferris research states that most spam falls into the follow-ing categories:

- Fake pharmaceuticals
- Fake fashion items (for example, watches)
- Pornography and prostitution
- Stock kiting—that is, spammers driving up the price of stocks by inciting victims to

buy them (also known as "pump and dump")

- Phishing and other fraud, such as "Nigerian 419" and "Spanish Prisoner"
- Trojan horses attempting to infect your PC with malware
- Misdirected non delivery reports and auto replies sent by badly configured mail servers replying to forged email ("backscatter")
- Spam from other types of senders, such as ignorant marketers, rogue affiliates, and misguided politicians or charities.

Spam filtering in Internet email can operate at two levels, an individual user level or an enterprise level. A person working at home is individual user and sending and receiving email via an ISP. Such a user who wishes to identify and filter spam email installs a spam filtering system on her individual PC. This system will either interface directly with their existing Mail user agent (MUA) (more generally known as the mail reader) or more typically will act as a MUA itself with full functionality for composing and receiving email and for managing mailboxes. Mails are filtered at Enterprise-level spam filtering as it enters the internal network of an enterprise. The software is installed on the mail server and interacts with the Mail transfer agent (MTA) classifying messages as they are received .Spam email, which is identified by the enterprise spam filter, will be categorized as a spam message for all users on that network. At an individual level spam can be filtered Son a LAN also. A networked user can choose to filter spam locally as it is downloaded to their PC on the LAN by installing an appropriate system. The current spam filtering systems uses rule-based scoring techniques at vast majority. A message is applied by set of rules and a score accumulates based on the rules that are true for the message. Systems include hundreds of rules and these rules need to be updated regularly as spammers alter content and behaviour to avoid the filters. Systems also incorporate list-based techniques where messages from identified users or domains can be automatically blocked or allowed through the filter letter followed by a period.

II. LITERATURE SURVEY

Liu Yuguo[7] has presented status of SVM in spam filtering, analyzing the effect of kernel function in SVM. They have proposed a word sequence kernel based on the dependent measure (PDWSK) model and applied to the spam filtering. They have compared the result of PDWSK model with other SVM under different kernel functions model in terms of all performance measures which gives higher accuracy when kernel function considers more text information. SumitSahu et al. [6] have proposed Meta. Multiclass-Classifier techniques to spam Email classification with feature selection techniques. They have compared result of all models in terms of recall, precision and accuracy. The Meta multiclass-Classifier techniques outperforms with other models. IsmailaIdris [5] has proposed neural network model for spam email classification. They have compared the result of both neural network and SVM model in terms of accuracy. The accuracy of neural network is at 94.017% with false positive rate of 0.299% as the better model than SVM. Aman Kumar Sharma et al. [4] have presented the decision tree algorithms ID3, J48, AD tree and simple CART of spam filtering. They have compared the result of all these algorithms in terms of accuracy. In terms of error measure suggested by author, j48 gave accuracy of 92.77%.R. Parimala et al. [3] have suggested GP-SVM to classify the spam email classification. In this study the feature selection techniques applied with various models but accuracy is lesser than the GP-SVM model with 70% features. The suggested model gave performance measures in terms of accuracy with feature selection and compare this model with others models. The GP-SVM suggested by author out-performs other ensemble model of SVM with feature selection techniques. W. A. Awad et al. [2] have suggested Naïve Baye's classifier of spam email classification. They have compared the suggested model with other models such as SVM, KN, NN, AIS, and RS in terms of recall, precision, and accuracy. The suggested model gave 99.46% accuracy, 98.46% recall and 99.66% precision with feature selection. The Naïve Bayes classifier outperforms the other models.

III. VARIOUS SPAM CLASSIFICATION TECHNIQUES

A. Decision tree:

The Decision tree [8] is probably the most popular data Mining technique. The most common data mining task for a tree is classification .The principle idea of a decision tree is to split our data recursively into subsets so that each subset contains more or less homogeneous states of your target variable (predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When this recursive process is completed, a decision tree is formed. There are various types of decision tree like CART,QUEST,CHAID,C4.5 etc.

B. Support Vector Machine (SVM):

Support vector machines (SVMs) [11] are supervised learning methods that generate input-output mapping functions from a set of labelled training data. The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., li-nearly separable) compared to the original input space. Then, the maximum-margin hyper planes are constructed to optimally separate the classes in the training data. Two parallel hyper planes are constructed on each side of the hyper plane that separates the data by maximizing the distance between the two parallel hyper planes. The larger the margin or distance between these parallel hyper planes the better the generalization error of the classifier will be made as a assumption.

C. Neural Network:

A neural network [10] contains a set of nodes (neurons) and edges that form a network. There are three types of nodes: input, hidden, and output. Each edge links two nodes with an associated weight. The direction of an edge represents the data flow during the prediction process. Each node is a unit of processing. Input nodes form the first layer of the network. In most neural networks, each input node is mapped to one input attribute such as age, gender or income. The original value of an input attribute needs to be massaged to a floating number in the same scale (often between –1 to 1) before processing.

D. Bayesian Classification:

Bayesian classifiers [9] are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Baye's theorem. Classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

.

V. CONCLUSIONS

Designing and developing a robust classifier for E-mail data classification in order to provide security to the E-mail users is challenging task and is a major research area. Researchers are using many techniques to design and develop a suitable classifier for this purpose. In this study, various researchers have applied different data mining classification techniques, statistical technique for spam email classification. We observe that accuracy of algorithm depends on data set and number of features, therefore every algorithm has its advantages and disadvantages.

REFERENCES

[1]. MdRafiqul Islam et al., "An innovative analyzer for multi-classifier e-mail classification based on grey, list analysis", Journal of Network and Computer Applications , vol. 32 ,pp. 357–366,2009.W.-K. Chen, Linear Networks and Systems. Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)

[2] W. A. Awad, S.M. Elseuofi, "Machine Learning methods for email classification", International Journal of Computer Applications,Vol. 16, pp. 39-45, 2011.K. Elissa, "An Overview of Decision Theory," unpublished. (Unplublished manuscript)

[3] R. Parimala, R. Nallaswamy, "A Study of Spam E-mail classification using Feature selection package",Global Journal of Computer Science and Technology Volume 11 Issue 7 ,Version 1.0, 2011.C. J. Kaufman, Rocky Mountain Research Laboratories, Boulder, Colo., personal communication, 1992. (Personal communication)

[4] Aman Kumar Sharma and Suresh Sahan, "A comparative study of classification algorithms for spams Email data Analysis", International Journal on Computer Science and Engineering (IJCSE),Vol. 3,pp. 1890-1895, 2011.S.P. Bingulac, "On the Compatibility of Adaptive Controllers," Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16, 1994.

[5] IsmailaIdris, "E-mail spam classification with ANN and Negative selection algorithms", International Journal of Computer Science & Communication Networks, Vol 1, pp.227-231, 2011.

[6] SumitSahu, BhartiDongre, Rajesh Vadhwani, “Web Spam Detection Using DifferentFeatures”, International Journal of Soft Computing and Engineering (IJSCE)ISSN: vol. 1,pp. 2231-2307, 2011

[7] Liu Yuguo , Zhu Zhenfang , Zhao Jing, “A word sequence kernels used in spam-filtering”,Scientific Research and Essays Vol. 6 , pp. 1275-1280, 2011.

[8] ZhaoHui Tang, Jamie Maclennan, “Data Mining with SQL Server 2005”,Willey ublishing,Inc,USA,2005

[9] Jaiwei Han and MichelineKamber, “Data Mining Concepts and Techniques “, Morgan Kaufmann, San Francisco, Second Edition 2006.

[10] .ZhaoHui Tang, Jamie Maclennan, “Data Mining with SQL Server 2005”, Willey ublishing, Inc, USA, 2005.

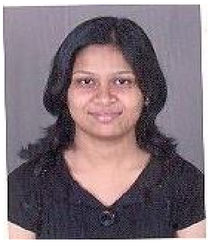
[11]. Dr. David L. Olson, Dr. DursunDelen,”Advanced Data Mining Techniques “, New York, 2008

Authors:

[1] Aawes h kumar Borkar

He received his B.E. degree in Computer Science and Engineering from Govt. Engineering College, Bilaspur under Guru Ghasi Das University, Bilaspur and M. Tech. pursuing with Computer Technology From RCET Bhilai under Swami Vivekananda Technical University, Bhilai.
His research area of interest is network security.

[2] Ms. Anshul Singh



She received her B.E. degree in Information Technology from CCET and M .Tech. Computer Science and Engineering from RCET Bhilai under Chhattisgarh Swami Vivekananda Technical University, Bhilai.
Presently working as Assistant professor at Department of Computer Science and Engineering RCET Bhilai, with an experience of 3 years of teaching .
She has published 7 papers in various International and National level. Her research area of interests is Data Mining and algorithms