International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 9, September - 2013

Spam Mail Filtering Through Data Mining Approach –A Comparative Performance Analysis

Yukti Kesharwani,
M.Tech. Scholor
R.K.D.F College
Mandideep Road Bhopal,(M.P.)

Shrikant Lade
HOD of IT Deptt.
R.K.D.F. College
Mandideep Road Bhopal,(M.P.)

Abstract

Spam is also known junk mail. E mail is undoubtedly a very effective, cheap and easy method of communication these days. But, due to social networks and advertisers, most of the emails contain unwanted information called spam. Generally, a spam is e-mail advertising for some product sent to a mailing list or newsgroup. In order to filter the messages and separate the genuine messages from the junk mail, the spam filters are preferred. Even though more number of classification techniques has been developed for spam classification, still none of the algorithms produces 100% accuracy, In this paper, spam dataset is analyzed using CLEMENTINE data mining tool for email spam classification. Initially, various classification algorithms are applied over this dataset and cross validation is done for each of these classifiers. Finally, best classifier for email spam is identified based on the Training and testing accuracy of various models and Performance measures.

Keywords: **classifier, e-mail , spam , spam filters, datamining tool .**

I. INTRODUCTION

Internet has become an indispensable method to communicate with each other,because of its fast,effective and cheap communication way. This enables internet user to easily transfer information from anywhere in the world in a fraction of second increases and it saves a lot of time and cost.Therefore spammers prefer to send spam through such kind of communication. The threat of unsolicited junk emails, known as spams, becomes more and more serious. The volume of junk mail has grown enormously in the past few years and consequently they are faced with spam problem. E-mail Spam is non requested information sent to the E-mail boxes and such a big problem for users. Spammers collect e-mail addresses from,websites, chatrooms, customer lists etc are the other source from they collect and are sold to other spammers. Spam can contain malware, when user open the spam the malware silently get installed on the system. The spam can also affect the financial information like bank account number, password and other private information. Email spam has steadily grown since the early 1990s. Botnets, networks of virus-infected computers, are used to send about 80% of spam. Since the expense of the spam is borne mostly by the recipient, it is effectively postage due advertising.Therefore, email classification becomes an important research area to automatically classify original emails from spam emails. The blind posting of unsolicited email messages, known as spam, is an example of the misuse . A spam filter is an email service feature designed to detect unsolicited and unwanted email and prevent those messages from a user's inbox. Because a large amount of global email messages are spam, effective spam filters are critical to maintaining clean and spam-free inboxes. More number of classification techniques has been developed for spam classification, still 100% accuracy of predicting the spam email is questionable.

This paper focuses on the classification of spam E-mails using **CLEMENTINE** data mining techniques and **Spam-Email data** set collected from **UCI repository.** Our purpose is not only to filter messages into spam and not spam, but still to divide spam messages into thematically similar groups and to analyze them, in order to define the social networks of spammers.

II. RELATED WORK

The e-mail spam problem is increasingly serious nowadays. These spam mails have already caused many problems such as time and effort must be devoted to either deleting it after it is received, or preventing it from even reaching the user[8],filling mailboxes,wasting network bandwidth.various techniques have been explored to relieve the problem and the one of the most important technique is filtering. The task of spam filtering is to filter out unsolicited e-mails atomatically from a user's mail stream. Most analyses of spam have focused exclusively on

email [1, 11, 12]. They characterize spam from various aspects such as the behavior of spammers [11, 12]. Paper [15] formalizes a problem of clustering of spam message collection through criterion function. Rambow et al. apply a machine learning approach to email summarization [3]. The machine learning approach consists of the automatic construction of a classifier based on a training set [14-16]. Approaches to filtering junk email are considered [9,10,6,7] showed approaches to filtering emails involve the deployment of data mining techniques.The authors Lixin Fu and Geetha Gali [17], have worked with Bayesian algorithm to filter e-mail spams. An anti-spam filtering technique was presented by [5]; His techniques are centered on artificial neural network (ANN) and Bayesian Networks.
Sakkis et al. [13] combined a Naïve Bayes (NB) and k-nearest neighbor (k-NN) classifiers by stacking method and found that the ensemble achieved better performance. In paper [2,4], automatic antispam filtering becomes an important member of an emerging family of junk-filtering tools for the Internet,which will include tools to remove advertisements.

III. PROPOSED SYSTEM

The work deals with this paper, e-mail spam detection and blocking of spam messages for spammers.In this section, we propose a new method of filtering spam based on preserving the sequence of term occurrence in datamining techniques. we present the data sets used in our analysis and discuss how we identify spam messages. In the context of spam filtering, a number of ensemble classification methods have been studied. Subsequently, we present the corresponding classification model. The overall design of the proposed system is given in Fig. 1 and each of these components is addressed in the following sections briefly. This proposed system is made up of four components are spam data set, classification model, Trained classification model & classification result.

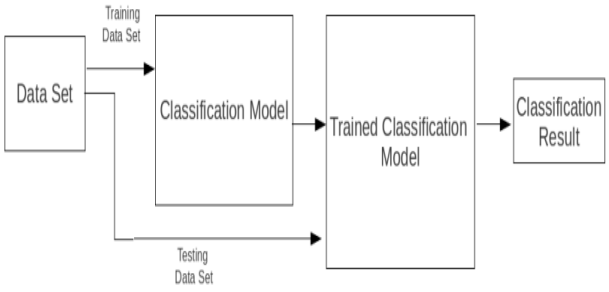

Fig 1. Architectural design of the proposed system

Spam Data Sets

Spam-Email dataset are collected from **UCI repository**.This dataset contains spam amd legitimate message and also,suitable for use in testing spam filtering system. This dataset contains 4601 instances and 58 attributes (57continuous input attribute and 1 nominal class label target attribute.

Training and Testing Data sets

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, we can minimize the effects of data discrepancies and better understand the characteristics of the model.
After a model has been processed by using the training set, we test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that we want to predict, it is easy to determine whether the model's guesses are correct.

IV. CLASSIFICATION MODEL

SVM

Support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

CHAID

CHAID is a type of decision tree technique, based upon adjusted significance.The technique was developed in South Africa and was published in 1980 by Gordon V. Kass, who had completed a PhD thesis on this topic. CHAID can be used for of CHAID being originally known as XAID) as well as prediction (in a similar fashion to regression analysis, this version classification, and for detection of interaction between variables. CHAID stands for **CH**i-squared **A**utomatic **I**nteraction **D**etection, based upon a formal extension of the US AID (Automatic Interaction Detection) and THAID (THeta Automatic Interaction Detection) procedures of the 1960s and 70s, which in turn were extensions of earlier research, including that performed in the UK in the 1950s.

In practice, CHAID is often used in the context of direct marketing to select groups of consumers and predict how their responses to some variables affect other variables, although other early applications were in the field of medical and psychiatric research.

Like other decision trees, CHAID's advantages are that its output is highly visual and easy to interpret. Because it uses multiway splits by default, it needs rather large sample sizes to work effectively, since with small sample sizes the respondent groups can quickly become too small for reliable analysis.

ARTIFICIAL NEURAL NETWORK

Artificil neural network are models inspired by animal central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network.

For example, in a neural network for handwriting recognition, a set of input neurons may be activated by the pixels of an input image representing a letter or digit. The activations of these neurons are then passed on, weighted and transformed by some function determined by the network's designer, to other neurons, etc., until finally an output neuron is activated that determines which character was read.

Like other machine learning methods, neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

CHART

The CART algorithm is based on Classification and Regression Trees by Breiman et al (1984). A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample.

QUEST

QUEST is proposed by Loh and Shih (1997), and stands for Quick, Unbiased, Efficient, Statistical Tree. It is a tree-structured classification algorithm that yields a binary decision tree. A comparison study of QUEST and other algorithms was conducted by Lim et al (2000). The QUEST tree growing process consists of the selection of a split predictor, selection of a split point for the selected predictor, and stopping. In this algorithm, only univariate splits are considered.

V. CONCLUSION

In this proposed work a frame work was established for new approach in spam detection . In this paper, we experiment several data sets and Separating data into training and testing sets. Typically, when we separate a data set into a training set and testing set, most of the data is used for training and a smaller portion of the data is used for testing. After a model has been processed by using the training set, we test the models by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that we want to predict, it is easy to determine whether the model's guesses are correct.This method helps to identify the unwanted information and threats. Further extended this work in a way of comparision of different classification algorithm with another classfication algorithm and provide the different output which provided by them and finding accuracy and precision of them.

REFERENCES

[1] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker. Spamscatter: Characterizing internet scam hosting infrastructure. In Usenix Security, 2007.
[2] Androutsopoulos .I, J. Koutsias, K.V. Chandrinos, G. Paliouras, and C.D. Spyropoulos. An Evaluation of Naive Bayesian Anti-Spam Filtering. Proceedings of the Workshopon Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain, pages 9–17, 2000.
[3] N. Ducheneaut and V. Bellotti. E-mail as habitat: an exploration of embedded personal information management. Interactions v.8, n.5, pp.30-38, 2001.
[4] Apte, C. and F. Damerau. Automated Learning of Decision Rules for Text Categorization. ACM Transactions on Information Systems, 12(3):233–251, 1994.
[5] L. Özgür, *et al.*, "Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish," *Pattern Recognition Letters*, vol. 25, pp. 1819-1831, 2004.

[6] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," In Proc. Of the AAAI Workshop on Learning for Text Categorization.
[7] T. Fawcett, "in vivo spam filtering: A challenge problem for data mining," In Proc. of ninth KDD Explorations vol.5 no.2, 2003.
[8]. Message Labs Spam Intercepts data, 2006, http://www.messagelabs.com/publishedcontent/publish/threat_watch_dotcom_en/threat_statistics/spam_intercepts/DA_114633.chp. htm
[9] W. Cohen, "Learning rules that classify e-mail," In Proc. of the AAAI Spring Symposium on Machine Learning in Information Access, 1996.
[10] Y. Diao, H. Lu, and D. Wu, "A comparative study of classification based personal e-mail filtering," In Proc. Of fourth PAKDD, 2000.
[11] A. Ramachandran and N. Feamster. Understanding
Sigcomm, 2006.
[12] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In ACM CCS, 2007.
[13] Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos. Stacking classifiers for anti-spam filtering of e-mail. Proceedings of 6th Conference on Empirical Methods in Natural Language Processing, 1:44-50, 2001.
[14] J. Wang, K. Gao, H.Q. Vu, "SpamCooling: A Parallel Heterogeneous Ensemble Spam Filtering System Based on Active Learning Techniques, Journal of Convergence Information Technology, vol. 5, no. 4, pp. 90-102, 2010.
[15] Perkins, A. The classification of search engine spam. http://www. ebrand management.Com/ white papers/spam classification.
[16] W. Hsu, T. Yu, "E-mail Spam Filtering Based on Support Vector Machines with Taguchi Method for Parameter Selection", Journal of Convergence Information Technology, vol. 5, no. 8, pp. 78- 88, 2010.
[17] Lixin Fu and Geetha Gali, "Classification Algorithm for Filtering Email spam", Recent Progress in Data Engineering and Internet Technology, 2012, Volume 157, 149-154, DOI: 10.1007/978-3-642-28798-5_21.