Email Spam Classification Based on Supervised Learning Algorithms

K.Uma Shankar[1],G.Jaya Bharathi[2]
Assistant Professor, Department of Computer Science and Engineering,
K.G Reddy College of Engineering and Technology. Hyderabad, Telangana.

Abstract: In data mining, e-mail spam is a serious threat to business and industry. Reducing spam and preventing the accumulation of spam stored in the user's mailbox is a challenge for users. Identifying the best algorithm for classifying spam is an important task. In this context, we use decision tree algorithms to filter spam, because the primary job is to organize spam or ham mail. The algorithms are created; filtering algorithms previously tested is applied. The results of the different algorithms are evaluated from the accuracy, the error rate, the accuracy and the actual speed of the error. Comparing the above algorithms based on their performance shows that our proposed algorithm performs better than other classifiers before and after applying weka filters.

Keywords: e-mail spam, decision tree algorithm, data mining, and classification.

1. Introduction

E-mail is an efficient, quick and low-cost communication approach. E-mail Spam is no requested data sent to the E-mail boxes. Spam could be a huge drawback each for users and for ISPs.According to investigation nowadays user receives a lot of spam emails then non spam emails. To avoid spam/irrelevant mails we'd like effective spam filtering strategies. Spam mails area unit used for spreading virus or malicious code, for fraud in banking, for phishing, and for advertising. Therefore it will cause major problem for web users like loading traffic on the network, wasting looking out time of user and energy of the user, and wastage of network information measure etc. There are several approaches are used for spam email classification consistent with data outside of the content of email messages those completely different e-mail classification techniques are together with rule based mostly and content based techniques.

Email Filtering/Spam Filtering:

To detect unsolicited and unwanted email and prevent those unwanted messages from getting to a user’s inbox is called spam filter. The spam filter is a program like other types of filtering program looks for certain criteria on which it bases judgments. The input of email filtering software is emails. The message through unchanged for delivery to the user's mailbox is the output of email filter. Some of the mail filters are able to edit messages during processing. Mail filters have differing degrees of configurability. Once in a while they settle on choices taking into account coordinating a consistent expression. Different times,

essential words in the message body are utilized, or maybe the email location of the sender of the message. Some more propelled channels, especially hostile to spam channels, use measurable archive order methods, for example, the guileless Bayes classifier. Picture sifting can likewise be utilized that utilization complex picture examination calculations to identify skin-tones and particular body shapes typically connected with obscene pictures. Mail filters can be introduced by the client, either as independent projects (see interfaces underneath), or as a major aspect of their email project (email customer).

Several machine learning algorithms have been used in spam e-mail filtering, but in Supervised Learning Algorithms are particularly popular in commercial and open-source spam filters [2]. This is because of its simplicity, which make them easy to implement and just need short training time or fast evaluation to filter email spam. The filter requires training that can be provided by a previous set of spam and non-spam messages. It keeps track of each word that occurs only in spam, in non-spam messages, and in both. The several different methods to identify incoming messages as spam are, Whitelist / Blacklist, Bayesian analysis, Mail header analysis, Keyword checking, K nearest neighbors, Support vector machine (SVM), Neural Networks based spam filteration,or by technique of genetic engineering can also be applied for spam filter creation recently.

2. Related Work

Nowadays, e-mail provides many ways to send millions of advertisement at no cost to sender. As a result, many unsolicited bulk e-mail, also known as spam e-mail spread widely and become serious threat to not only the Internet but also to society. For example, when user received large amount of e-mail spam, the chance of the user forgot to read a non-spam message increase. As a result, many e-mail readers have to spend their time removing unwanted messages. E-mail spam also may cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content. Over the past many years, many approaches have been provided to block e-mail spam [1].

Web spam which is a major issue throughout today's web search tool; consequently it is important for web crawlers to have the capacity to detect web spam amid creeping. The Classification Models are designed by machine learning order algorithm. [2] The one machine learning algorithm is Naïve Bayesian Classifier which is also used in [1] to separate the spam and non-spam mails. Big Data analyzing framework which is also outline for spam detection. Extricating the feeling from a message is a method for get the valuable data. In Machine learning innovations can gain from the preparation datasets furthermore anticipate the choice making framework hence they are broadly utilized as a part of feeling order with the exceptionally precision of framework. [3]

Most of the research work has already been carried out on improving the efficiency and accuracy of Naïve Bayesian approach. Paul Graham’s Naïve Bayesian Machine learning approach is used to improve

the efficiency of Bayesian approach. [1] For vast dataset also using the naïve Bayesian algorithm and increment the precision of NBC. [4] The research work has also carried out for increase the accuracy and time efficiency of system.

Several machine learning algorithms have been employed in anti-spam e-mail spam filtering, including algorithms that are considered top-performers in Text Classification [5], like Boosting algorithm, Support Vector Machines (SVM) algorithm [6] and Na¨ıve Bayes algorithm [7]. Konstantin Tretyakov et al., [6] have evaluated several most popular machine learning methods i.e., Bayesian classification, k-NN, ANNs, SVMs and of their applicability to the problem of spam-filtering. In this work, the author proposed most trivial sample implementation of the named techniques and the comparison of their performance on the PU1 spam corpus dataset is presented. The author used extracting feature to convert all messages to vectors of numbers (feature vectors) and then classify these vectors. This is because most of the machine learning algorithms can only classify numerical objects like vector. Then the author created the straightforward C++ implementations of the algorithms, and tested them on the PU1 spam corpus taken from http://www.stat.purdue.edu/~mdw/598/datasets.html. The PU1 corpus consists of 1100 messages, of which 489 are spam. The test setup use by efficiency measure which are precision, legitimate mail fallout and spam fallout. From the result, the performance of the k-nearest neighbors classifier appeared to be poor and the number of false positives was always rather large. According to the author, only the Decision Tree algorithm has passed the test.

3. System Study

Presented System

Email Spam is most crucial matter in a social network. There are many problem created through spam. The spam is nothing this is unwanted message or mail which the end user doesn't want in our mail box. Because of these spam the performance of the system can be degraded and also affected the accuracy of the system. To send the unsolicited or unwanted messages which are also called spam is used in Electronic spamming. In this paper we presented regards the email spam, where how spam can spoil the performance of mailing system. In the previous study there are many types of spam classifier are present too detect the spam and non-spam mails.

There are different email filtering techniques are also used in spam detection. Mostly popular filters or classifier are: Decision tree classifier, Negative Selection Algorithm, Genetic Algorithm Support Vector Machine Classifier, Bayesian Classifier etc. From the previous study we identify that Support Vector Machine (SVM Classifier) are used for email spam classification. But it takes very much time for detecting spam. The SVM Classifier has also wrongly classified the messages. So the system can be on a risk. The error rate of SVM Classifier is very high. In this paper there is also discussion in the Feature Selection

process. There are different feature extractions techniques are present which are used in extracting the messages.

Proposed system

To solve the problem of previous study, we proposed the decision tree classifier for classify the spam and non-spam mails. The decision tree classifier is one of the most popular and simplest methods for classification. Decision tree classifier are highly scalable, learning problem the number of features are required for the number of linear parameter. Training of the large data simple can be easily done with decision tree classifier, which takes a very less time as compared to other classifier. The accuracy of system is increase using decision tree classifier**.**

Algorithm Used

Decision Tree

In this paper we have taken different decision tree classifiers and apart from other types of data mining classifiers, we emphasized specifically on decision tree classifiers for the particular application of spam filtration technique. This is done because of decision tree filters are easy to implement and easy to understand. It provides an overall satisfactory performance as far as spam mail detection is concerned. Decision tree learning is a method commonly used in data mining. The goal is to create a DT model and train the model so that it can predicts the value of a target variable based on several input variables. An example is shown on the below. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A tree can be "learned" by splitting the source set into various subsets based on attribute value prefixed. This process is repeated on each derived subset in a recursive manner which is called as recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. An example of a simple decision tree is shown in Fig.1 corresponding to the training data set given in Table1.The leaf nodes represent the decision of buying computer for different people with different criteria[9].

In data mining, decision trees can also be described as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data. Data comes in records of the form

(x , Y) = (x1, x2, x3, …., xk, Y) The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the input variables, x1, x2, x3…etc., that are used for that task.

The decision tree generated by C4.5 can be used for various classification problems. At each node of the tree the algorithm chooses an attribute that can further split the samples into subsets. Each leaf node represents a classification or decision. Some premises guide this algorithm, such as the following [8]

If all cases are of the same class, the tree is a leaf and so the leaf is returned labeled with this class;

For each attribute, calculate the potential information provided by a test on the attribute (based on the probabilities of each case having a particular value for the attribute). Also calculate the gain in information that would result from a test on the attribute (based on the probabilities of each case with a particular value for the attribute being ofa particular class)

Depending on the current selection criterion, find the best attribute to branch on. J48 is an open source implementation of C4.5. Decision tree is built by analyzing data the nodes of which are used to evaluate significance of existing features.

Feature Selection

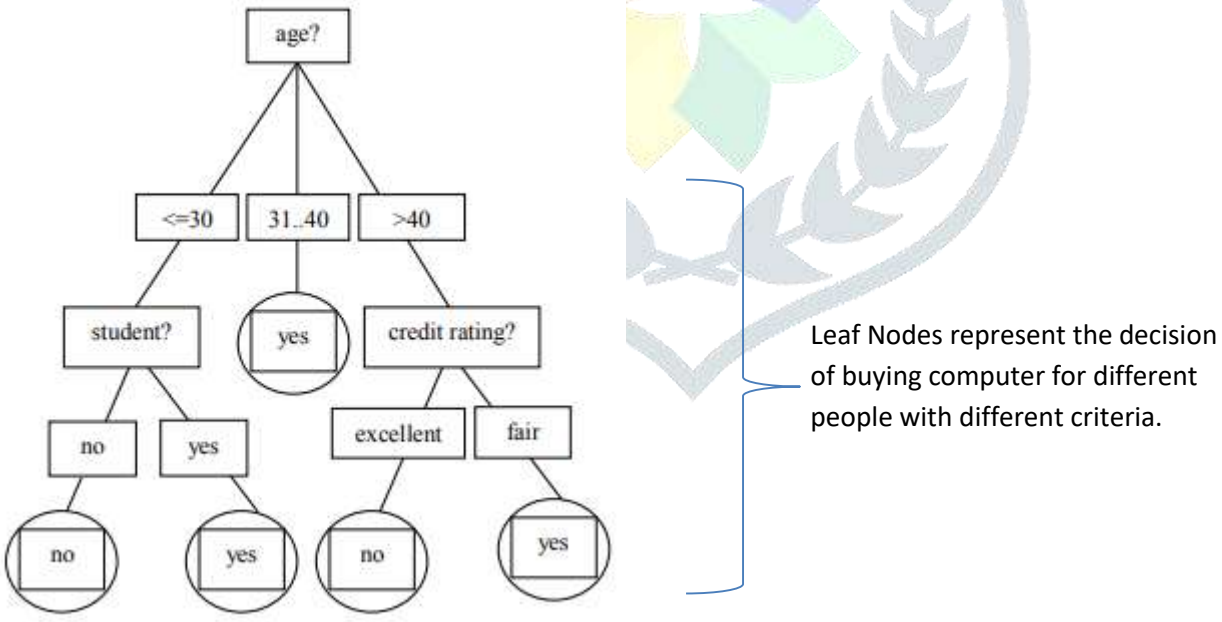

Figure 1. Sample decision tree corresponding dataset.

The dataset has been prepared to act for the feature selection after the removal of unnecessary stop words such as “The”, “In”, “A”, “On” etc. Now, moving on, we can say that the work is based on rules and uses a binary score-based system. The rules are framed by analyzing the mail header information, keyword

matching and the body of the message. The score assigned is 0 if the rule verifies to be false, else 1. There are number of rules framed by considering the various features that will aid to identify the spam messages effectively. The rules are present for each category of mail spam or ham. For example for the rule "From Correct Domain Name" if the feature corresponding to this rule is "www.way2sms.com" it will be considered as a spam feature and the score of 1 will be added to the composite spam-score which is nothing but the measure of spam strength for a mail. However, if the feature "From Correct Domain Name" is say, "www.wbut.net" then it would be considered as ham feature and 0 would be added to the score. Each rule performs a test on the email, and each rule has a score. When an email is processed, it is tested against each rule. For each rule found to be true for an email, the score associated with the rule is added to the overall score for that email. Once all the rules have been used, the total score for the email is compared to a threshold value [8]. If the score exceeds the threshold spam score value, then the email is marked as spam and the other mails are classified as legitimate ham mails. In this work, the rules used are:

Table 1. Scheme of Rules assigned to Spam Features

1	From Correct Domain Name
2	Blocked IP
3	Content Type
4	To header original
5	Is subject present
6	Is reply message
7	Is forwarded message
8	Sensual message
9	Subject content has vulgar words
10	Character set includes foreign language except English

4. Experimental Results

Weka, an open source, GUI based, portable workbench has been used to perform the analysis of various email spam filtering techniques with a rigorous data set applied. We created the data set of emails using attributes and relations from the spam mails received in the mailbox for over six months. There were 88 attributes and 256 instances taken as a total data set and 10 fold cross-validations has been done to test the result and compare the different results. The different decision tree algorithms we run using Weka are NBTree, C4.5 Decision tree classifier and Logistic Model Tree classifier and checked the performances with different criteria in terms of time , result efficiency and accuracy achieved by these various decision tree classifiers and some other criteria like false positive , false negative rates of decisions taken by these classifiers[10]. The performance has been measured using a number of parameters. We have used cross-validation for predictive accuracy. Training time shows how much time the classifier takes to show results

on the given data set. When we say incorrectly classified we mean that some mails have either been categorized as false positive or as false negative. False positive means ham has been classified as spam and false negative means spam has been classified as ham. Precision gives the percentage of total instances that have been correctly classified.

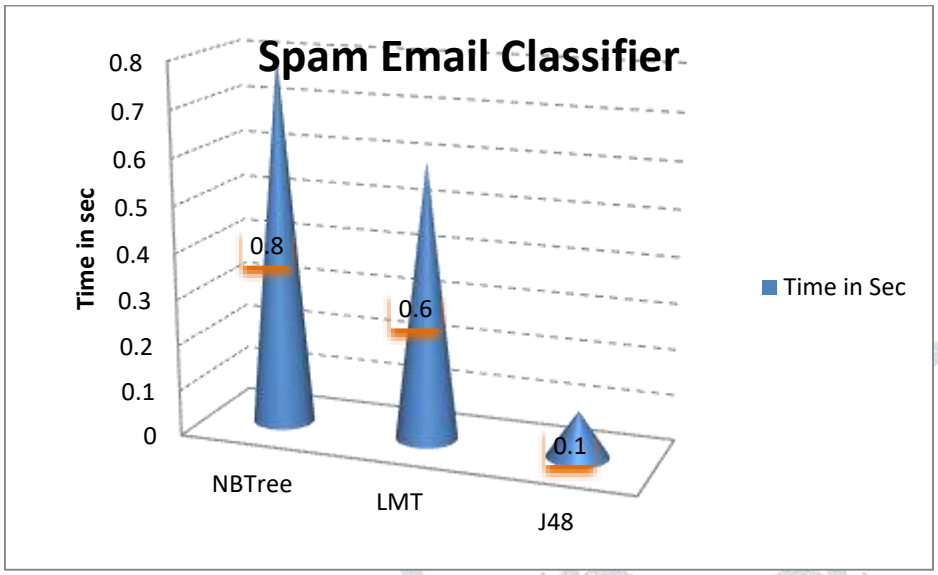

Figure 2. Email Spam Classifier time in Sec. with various Decision Tree classifiers

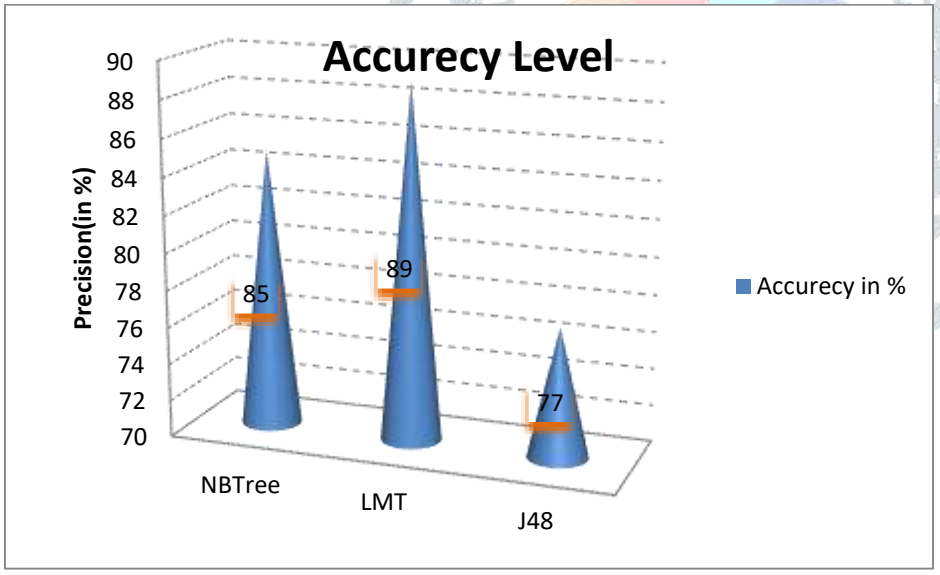

Figure .3 Accuracy level of the Decision Tree classifiers

5. Conclusion

Spam is a big threat to regular customers of emails, businesses, and several other sectors. In this paper, we have been compared the decision tree algorithm on three e-mails based processes that contain e-mails from other sources. Analysis of test data results shows the accuracy of LMT is high from NBT and J48 configuration. J48 is considered to be the best whenever training time is being considered as a critical

parameter because it takes minimum training time than other DT algorithms discussed here. Therefore considering overall performance we conclude that LMT classifier can be used safely for building reliable spam filters though J48 can be used for spam filter application but after properly justifying the scope of improvement in terms of false positive rate.

Reference

[1] Sharma K. and Jatana N. (2014)"Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach" IEEE 2014 pp. 939-942.

[2] Sharma A. and Anchal (2014), "SMS Spam Detection Using Neural Network Classifier",ISSN: 2277 128X Volume 4, Issue 6, June 2014, pp. 240-244.

[3] Ali M. et al (2014), , "Multiple Classifications for Detecting Spam email by Novel Consultation Algorithm", CCECE 2014, IEEE 2014, pp. 1-5.

[4] Liu B. et al (2013) "Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier" IEEE 2013 pp.99-104.

[5] Rathi, M. and Pareek, V. "Spam Mail Detection through Data Mining A Comparative Performance Analysis", I.J. Modern Education and Computer Science, 2013, 12, 31-39.

[6] Kumar, S., Gao, X., Welch, I. and Mansoori, M., "A Machine Learning Based Web Spam Filtering Approach", IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana, 2016, pp. 973-980.

[7] Tariq, M., B., Jameel A. Tariq, Q., Jan, R. Nisar, A. S., "Detecting Threat E-mails using Bayesian Approach", IJSDIA International Journal of Secure Digital Information Age, Vol. 1. No. 2, December 2009.

[8] Feng, W., Sun, J., Zhang, L., Cao, C. and Yang,Q., "A support vector machine based naive Bayes algorithm for spam filtering," 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC), Las Vegas, NV, 2016, pp. 1-8.

[9] V. Christina et al. Email Spam Filtering using Supervised Machine Learning Techniques. International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 09, 2010, 3126-3129.

[10]. Sarit Chakraborty, Bikromadittya Mondal. Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis, International Journal of Computer Applications (0975 – 888) Volume 47– No.16, June 2012