Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis

Sarit Chakraborty[1] Bikromadittya Mondal[2]
Department of Computer Science & Engineering,
B. P. Poddar Institute of Management & Technology, Kolkata, India

ABSTRACT
In recent years the highestdegree of communication happens through e-mails which are often affected by passive or active attacks. Effective spam filtering measures are the timely requirement to handle such attacks. Many efficient spam filters are available now-a-days with different degrees of performance and usually the accuracy level varies between 60-80% on an average. But most of the filtering techniques are unable to handle frequent changing scenario of spam mails adopted by the spammers over the time. Therefore improved spam control algorithms or enhancing the efficiency of various existing data mining algorithms to its fullest extent are the utmost requirement.In this paper three types of decision tree classifying techniques which are basically data mining classifiers namely Naïve Bayes Tree classifier (NBT), C 4.5 (or J48) decision tree classifier and Logistic Model Tree classifier (LMT) are studied and analyzed for spam mail filtration. The test results depict that LMT is giving the most efficient result in terms of performance with almost 90% accuracy level to detect spam mails and non-spam (HAM) mails.

Keywords
Spam, Ham,Data Mining,Naïve Bayes Tree, Logistic Model Tree, Machinelearning, Spam-score

1. INTRODUCTION
The amount of data kept in form of digital files and databases is growing at a phenomenal rate over the last couple of decades. At the same time the users of these data are expecting to find more and more sophisticated information and hidden data patterns from them. Amarketing manager is no longer satisfied with just customer contacts but also wants detailedinformation on them such as their personal details, past interests, past buying patterns etc. In simple words data in today's world is explosive, expanding, ever changing but not exhaustive in nature. Data mining steps and techniques are developed to solve these needs. Data mining is often defined as the technique for finding hidden information in a database.

Now a day, 80% of the data are stored in textual form – such as magazines, newspapers, documents, journals, emails, etc. To uncover the hidden information from such text data a varied form of data mining is used –called as textmining. Text mining can be broadly defined as a knowledge intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. A major application of text mining in today's world is in web text mining which is used here in the field of SPAM FILTERING.

Spam is briefly defined as ''unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient'' (Cormack and Lynam, 2005b). According to annual reports, the amount of spam is frightfully increasing. In absolute numbers, the average of spams sent per day increased from 2.4 billion in 2002 to 300 billion in 2010 [2].

So far many methods have been proposed to automatically classify messages as spam's or legitimates. Among all proposed techniques, machine learning algorithms have achieved more success [3].

Several methods or approaches that are considered top-performers in text categorization and extensively analyzed till date are support vector machines (SVM) and the well-known Naive Bayes classifier. But in this paper we have shown that apart from those popular methods, the Decision Tree classifiers also provides great result as far as spam detection is concerned and out of those DT classifiers Logistic Model Tree (LMT) classifier provides around 90% of performance accuracy which is almost or better than the Naïve Bayes and many other spam detection techniques till date analyzed.

2. SPAM FILTERINGTECHNIQUES
Spam filters can work at various levels from operational point of view but in general depending on the usability it works on two levels i.e. at the server level and at the user level or individual client level. Server level spam filters operate on universal rules which are same for all users. They filter mails as they arrive at the server depending on the rules defined beforehand which is nothing but supervised learning methods are applied majorly also known as knowledge engineering[4].User level spam filters filter mails that come from the network mail server andthey work on the each user's / clients individual terminal. Developments in the field of spam filtering so far can be broadly classified into two fields: those based on machine learning (ML) principle and those not based on machine learning principle (Carpinter, Ray 2006).

ML based filtering techniques can again be classified into complementary and complete solutions. The machine learning approach does not require specifying any rules explicitly rather a set of pre-classified documents is needed which is used as training samples. Aspecific algorithm is then used to learn the classification rules from this training data set.

Non-ML techniques comprise of heuristic analysis, signature based techniques, blacklisting and traffic analysis.

In heuristic approach, regular expressions are used to check for common spam phrases and characteristics. A much simpler way is the blacklisting technique in which a block list is created. It consists of spam words in form of vulgar words, porn related text contents, abusive texts etc. Apart from that senders' name, email addresses, subject titles, and contents of "To", cc, bcc, "From" field etc that are also considered to identify spam. Whitelists perform the opposite function. These can be implemented both at the server and user level. A white list is a list, which includes all addresses from which the users always wish to receive mail [5].

User can add email addresses or entire domains, or functional domains. An interesting option is an automatic whitelist management tool that eliminates the need for administrators to manually input approved addresses on the whitelist and ensures that mail from particular senders or domains are never flagged as spam.

Spam filters can be implemented at all layers, firewallsexist in front of email server or at MTA (Mail TransferAgent). Email Server to provide an integrated Anti-Spam and Anti-Virus solution offering complete email protection at the network perimeter level, before unwanted or potentially dangerous email reaches the network. At MDA (Mail Delivery Agent) level also spam filters can be installed as a service to all of their customers. And finally Email client user can have personalized spam filters that automatically filter mail according to the chosen criteria.

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 Networksbased spam filteration,or by technique of genetic engineering can also be applied for spam filter creation recently [6].

3. METHODOLOGIES

In this paper we have taken various 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.

Table1: Training data set

age	income	student	credit_ rating	buys_ computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no

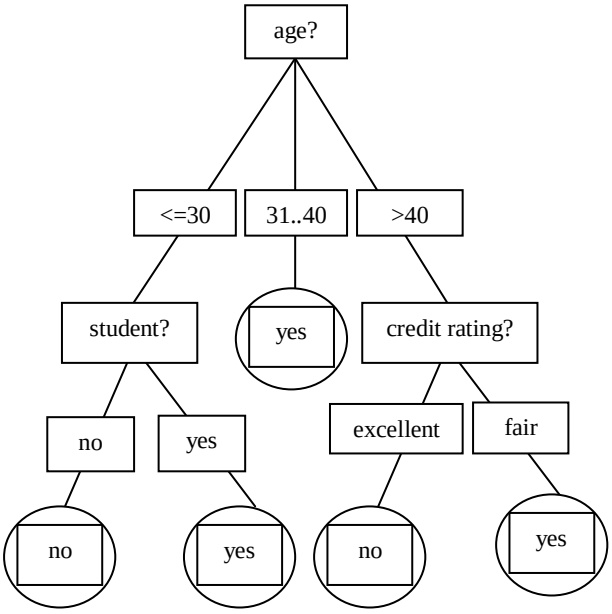

Fig.1 Decision Tree

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:

$$(x , Y) = (x_1, x_2, x_3, \ldots, x_k, Y)$$

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, x_1, x_2, x_3…etc., that are used for that task.

3.1 NBTree Classifier

This is a hybrid approach which combines the advantages of Naïve Bayes classifier and decision tree. In this classifier Naïve Bayes is applied at the nodes and decision tree is built with univariate splits at each node (Kohavi 2011).

In case of a large database NBTree classifier becomes very useful or if the database is of arbitrary size and the attributes are not necessarily independent. When we check spam mails, the database follows the above criterion. This classifier is as easy to interpret as Naïve Bayes and decision trees. The decision tree segments the data and each segment i.e. leafare described by Naïve Bayes.

3.2 C 4.5 / J48 Decision Tree Algorithm

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 labelled 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.

3.3 Logistic Model Tree Induction

A model tree consists of decision tree with logistic regression models at the leaves. Logistic Model Trees have been shown to be very accurate and compact classifiers in different research areas. Their greatest disadvantage is the computational complexity of inducing the logistic regression models in the tree. But the prediction of a model is obtained by sorting it down to a leaf and using the logistic prediction model associated with that leaf.A single logistic model is easier to interpret than C 4.5 trees. However building LMTs takes longer time. This can be shown by enough data and statistics. It can also be shown that trees generated by LMT are much smaller than those generated by C 4.5 induction. But in this paper it is shown that the training time required to build the Logistic model tree is lesser than the Naïve bayes classifier and interestingly correctly classifying ability for the particular application of spam mail filtration process as we have chosen is giving much better results in compare to Naïve Bayes classifier.

3.4 Feature Selection

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 2 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 105 attributes and 300 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 usingWeka 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.

The Table 3 below summarizes the experimental results. Comparative study of the performance has been graphically represented.

Table3 Comparative analysis of performance of the DT classifiers with Data Set 1

Evaluation Criteria / **Classifiers**	Training Time (in sec)	Correctly Classified Instances (Out of 300 Instances)	Precision /Accuracy (in %)	False positive (in %)
NBTree	42.06	245	81.66	22.8
J48	0.23	246	82.00	24.2
LMT	41.64	263	87.66	14.5

Here from Table 3 we first train and build the model with 300 numbers of instances with 105 numbers of attributes / spam words and then check with 10-fold cross-validation result. The training time to build the corresponding model is taken when it was run for the first time with a classifier and the precision value and False Positive value is taken later when we run that particular model / classifier with 10-fold cross validation.

But to ensure the consistency of the results and effectiveness of the discussed classifiers for different data set for the particular application of SPAM mail filteringwe took another small data set with 111 numbers of instances / Spam mails and 25 numbers of attributes and run all the respective classifiers on that data set once again. The results are given in Table 4But this time we run the classifiers with the respective model previously trained and build with data set 1.

Table 4 Comparative analysis of performance of the DT classifiers with Data Set 2

Evaluation Criteria / **Classifiers**	Training Time (in sec)	Correctly Classified Instances (Out of 111 Instances)	Precision /Accuracy (in %)	False positive (in %)
NBTree	0.91	92	82.88	33.3
J48	0.01	87	78.37	34.7
LMT	0.78	95	85.66	26.9

From Table 3 and Table 4 it can be inferred that both the data set is giving similar kind of results for the concerned classifiers of this paper. So now we can go for comparative analysis of the above mentioned classifiers and discuss their pros and cons in the light of various performance criteria.
As visible in Fig 3 and Fig 4, LMT outperforms NBTree and J48 when it comes to accuracy of results. LMT provides around 86% of accuracy and the False Positive rate is also much lower than NB Tree classifier and J48 classifier. NB Tree requires highest training time among all the decision tree classifiers discussed over here but the False Positive Rate i.e. chance of declaring a HAM mail as SPAM is in between J48 and LMT. The least value of False Positive rate can be achieved with LMT only which is a stringent requirement and quite logically needed criteria for real life filtering applications as no HAM mail supposed to be considered as SPAM though the reverse situation might be acceptable to some extent. Considering only training time as most desirable factor, J48 requires the least amount of training time and as well as least amount of run time among the same data set.

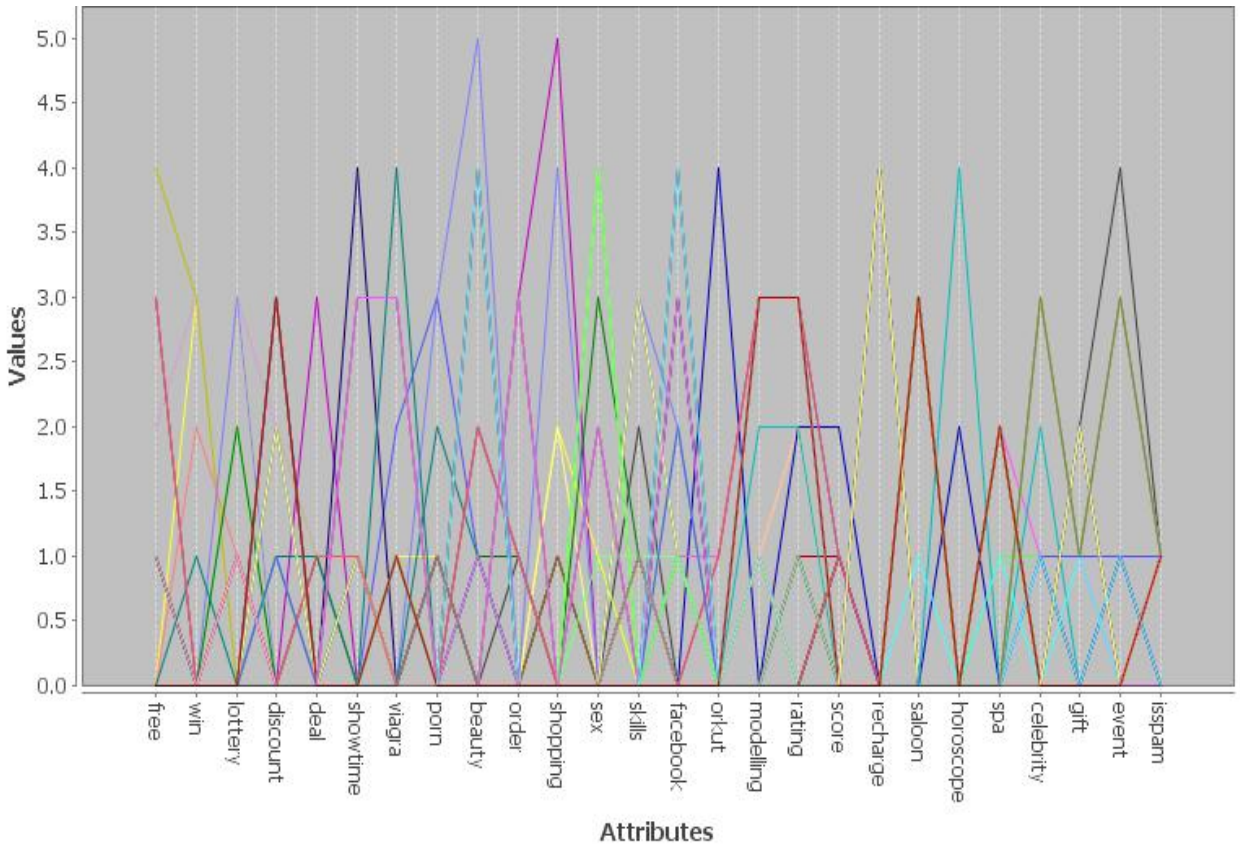

Fig.2Parallel Coordinates Plot: Values versus Spam- Attributes for Data Set 2

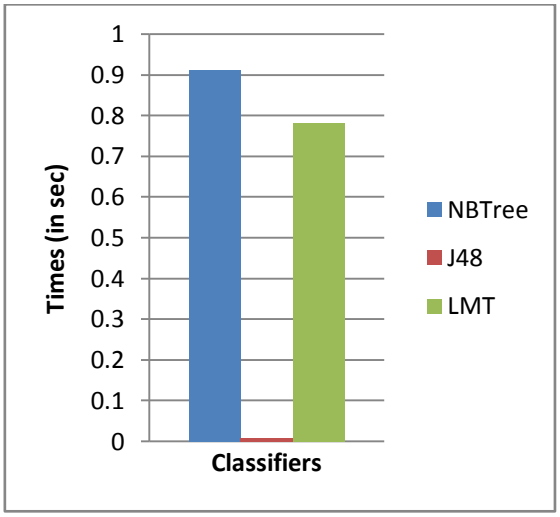

Fig.3 working time of various Decision Tree classifiers

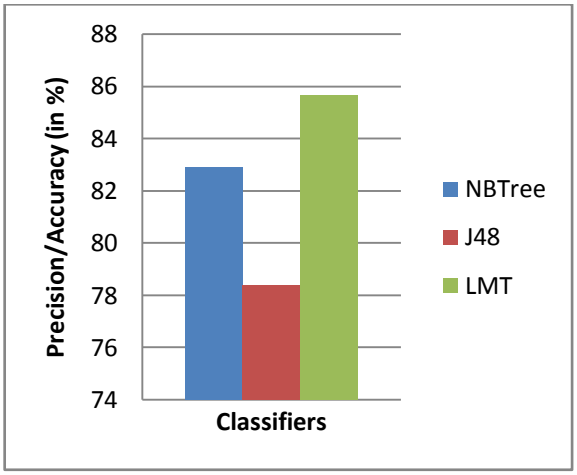

Fig.4Accuracy level ofthe Decision Tree classifiers

5. CONCLUSION

Spam is becoming a serious threat for regular users of emails, businesses and corporate firms. In this paper we have applied threedifferent decision tree algorithms on the set of emails which consist of spam and ham mails collected from our sources. Result analysis on test data showsthat LMT outperforms NBT and J48 classifiers when it comes to accuracy. J48 is considered to be the best whenever training time is being consideredas a critical parameter because it takes minimum training time than other DT algorithms discussed here. Therefore considering overall performance we concludethat 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.

6. REFERENCES

[1] P.Sudhakar, G.Poonkuzhali, K.Thaigarajan, K.Sarukesi, International Journal of Compuers, Issue 3, Volume 5, 2011, P. 332-345

[2] Almeida T, Yamakami A, Almeida J (2009) Evaluation of approaches for dimensionality reduction applied with Naive Bayes anti-spam filters. In: Proceedings of the 8th IEEE international conference on machine learning and applications, Miami, FL, USA, pp 517–522

[3] Cormack G (2008) Email spam filtering: a systematic review. Found Trends InfRetr 1(4):335–455

[4] Machine Learning Techniques in Spam FilteringKonstantin Tretyakov, kt@ut.ee, Institute of Computer Science, University of Tartu, Data Mining Problem-oriented Seminar, MTAT.03.177, May 2004, pp. 60-79.

[5] A Study on Email Spam Filtering Techniques, Christina V et. all. *International Journal of Computer Applications (0975 – 8887) Volume 12– No.1, December 2010, pp.07-09*

[6] Adaptive Spai Mail Filtering Using Genetic Algorithm,SanpakdeU et.all. Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference, 20-22 Feb. 2006, Vol 1, 441 - 445

[7] J.Quinlan. C 4. 5: Programs for Machine Learning. Morgan Kaufmann, 1992.

[8] 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

[9] Ahmed Khorsi, "An Overview of Content-based Spam Filtering Techniques", *Informatica*, vol. 31, no. 3, October 2007, pp 269-277.

[10] Weka. WEKA (Data Mining Software). Available athttp://www.cs.waikato.ac.nz/ml/weka/. 2006