iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** Issue I **Month of publication:** May 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Data Mining Approach on Various Classifiers in Email Spam Filtering

Mr. C. Balakumar [1], Dr. D. Ganeshkumar [2]
[1]PG Scholar, [2]Professor
Department of Computer Science and Engineering, P.A College of Engineering and Technology, Pollachi

Abstract— E-mails are the most nontrivial means of communication in the recent years. Spam mails often cause inconvenient to the users. The mails are classified as Spam and ham. Unwanted mails are called as spam and genuine mails are called as ham. In this paper, the effective decision tree classifiers are used to classify whether the mail is spam or ham. Many filtering techniques are used to find the spam mails and filter them but the accuracy and performance of the algorithms is distinct from each other. Efficient filtering of spam mails is an important requirement in using the existing data mining algorithms. In this paper, six decision tree algorithms that are basically used as classifiers namely J48 or C4.5, Rndtree, BFtree, REPtree, LMT and simple CART are compared. These algorithms were studied, analyzed and test results are shown in WEKA tool for efficient spam filtering. The results are compared and RndTree algorithm shows almost 99% accuracy level in filtering the spam mails and this shows best results among other classifiers.
Index Terms— classifiers, e-mail, ham, spam

I. INTRODUCTION

Spam is an unwanted usually commercial email sent to a large number of recipients. In internet spam has become an electronic thorn in the foot of the ubiquitous systems user. Spam can take away resources from users and service suppliers without compensation. [1]. Spammers collect e-mail addresses from group chats, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. Their content varies from deal to real estate to pornography. Since, the cost of the spam is borne mostly by the recipient, many individual and business people send bulk messages in the form of spam.In recent years, spam emails lands up into a serious security threat, and act as a prime medium for phishing of sensitive information. Addition to this, it also spread malicious software to various users. Therefore, email classification becomes an important research area to automatically classify original emails from spam emails. Spam email also fascinate problem for individuals and organizations because it is prone to misuse. Automatic email spam classification [4] contains more challenges because of unstructured information, more number of features and large number of documents. As the usage increases, all of these features may adversely affect performance in terms of quality and speed. Many recent algorithms use only relevant features for classification. Even though more number of classification techniques has been developed for spam classification, still 100% accuracy of predicting the spam email is questionable. So, identification of best spam algorithm itself became a tedious task because of features and drawbacks of every algorithm against each other.[2]. In this paper, spam dataset from UCI machine learning repository [3] is taken as input data for analyzing the various classification techniques using WEKA [5] data mining tool. In this work, feature selection is done first to select the relevant features for classification. After feature selection, six classification algorithms are taken for evaluation. In this evaluation process, different features are considered for choosing best spam filtering algorithm. Finally, performance evaluation is done to analyze the various classification algorithms to select the best classifier for spam emails.

II. RELATED WORKS

Email spam is one of the major problems of the today's Internet, bringing financial damage to companies and annoying individual users. Among the approaches developed to stop spam, filtering is the one of the most important technique. Spam mail, also called unsolicited bulk e-mail or junk mail that is sent to a group of recipients who have not requested it. The task of spam filtering is to rule out unsolicited e-mails automatically from a user's mail stream. These unsolicited mails have already caused many problems such as filling mailboxes, engulfing important personal mail, wasting network bandwidth, consuming users' time and energy to sort through it, not to mention all the other problems associated with spam [1].Developments in the field of spam filtering uses Machine Learning algorithms.

Machine learning algorithms are described as either 'supervised' or 'unsupervised'. The distinction is drawn from how the learner classifies data. In supervised algorithms, the classes are predetermined. These classes can be conceived of as a finite set, previously arrived at by a human. In practice, a certain segment of data will be labeled with these classifications. The machine

©IJRASET 2015: All Rights are Reserved

learner's task is to search for patterns and construct mathematical models. These models then are evaluated on the basis of their predictive capacity in relation to measures of variance in the data itself. Unsupervised learners are not provided with classifications. In fact, the basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. These groups are termed clusters.[6].

A. *What is a Spam Filter?*

A spam filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. For example, the simplest and earliest versions (such as the one available with Microsoft's Hotmail) can be set to watch for particular words in the subject line of messages and to exclude these from the user's inbox. This method is not especially effective, too often omitting perfectly legitimate messages (these are called *false positives*) and letting actual spam through. More sophisticated programs, such as Bayesian filters or other heuristic filters, attempt to identify spam through suspicious word patterns or word frequency.[7].

III. METHODLOGIES

Decision tree learning is a method commonly used in data mining. Decision Tree is a tree-structured plan of a set of attributes to test in order to predict the output. The example of decision tree is shown in Fig 1.This is the widely used learning method and it can be represented as If – Then rules.

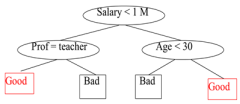


Fig 1- Decision Tree

In this paper, various decision tree classifiers are taken for evaluation and apart from other types of data mining classifiers are 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. The goal is to create a decision tree model and train the model so that it can predicts the value of a target variable based on several input variables. 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. *C4.5/J48 Decision Tree Algorithm*

C 4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. J48 is an open source Java implementation of the C 4.5 algorithm in the weka data mining tool. Time Sleuth extends C 4.5 use to temporal and causal discovery.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 in subsets. Every 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 of a 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.

The decision tree for J48 algorithm in WEKA tool is given in Fig 2.a, Fig 2.b, Fig 2.c

Volume 3, Special Issue-1, May 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

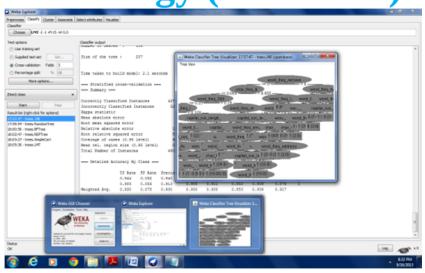


Fig.2.a J48 Classifier

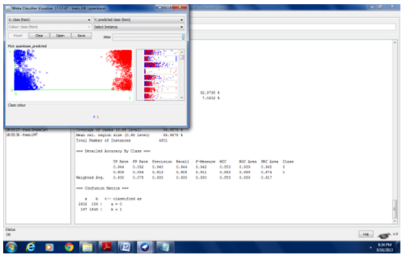


Fig.2.b Result of J48 (Classifier errors)

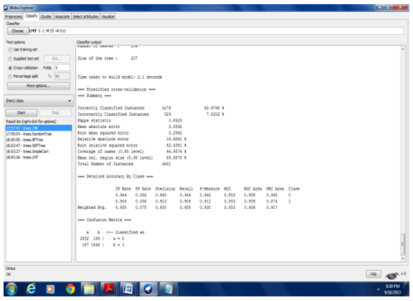


Fig.2.c Result of J48
(Results with confusion matrix)

B. Logistic Model Tree Induction

A model tree consists of decision tree with logistic regression models at the leaves. 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 C4.5 trees. However, building LMTs takes longer time.[9] 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 C4.5 induction. The execution time of this algorithm is given below.

Time taken to build model: 763.86 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	4313	93.7405 %
Incorrectly Classified Instances	288	6.2595 %
Kappa statistic	0.8687	
Mean absolute error	0.0828	
Root mean squared error	0.2334	

Relative absolute error 17.3369 %
Root relative squared error 47.7611 %
Coverage of cases (0.95 level) 97.3267 %
Mean rel. region size (0.95 level) 59.085 %
Total Number of Instances 4601

=== Confusion Matrix ===

```
   a    b   <-- classified as
 2653  135 |    a = 0
  153 1660 |    b = 1
```

C. SimpleCART Algorithm

This algorithm was first introduced by Breiman et al. The CART method under WEKA is a very popular classification tree learning algorithm. CART builds a decision tree by splitting the records at every node, according to the function of a single attribute it uses the gini index for determining the most excellent split. The CS-CRT is similar to CART but with cost sensitive classification.[10]

D. Random Forest Tree (Rnd Tree)

A Random Tree consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. For regression problems, the tree response is estimated for the dependent variable given by the predictors.[11].The accuracy level of this algorithm is very high and execution time is low compared to LMT which makes it to perform better than other algorithms. In this paper, it is shown that this algorithm is chosen for spam filtration process in giving much better results than the other classifiers.

E. REPTree

REPTree algorithm is a fast decision tree learner. It builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back-fitting). The algorithm only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).[13].

F. BFTree

This algorithm is a best-first decision tree classifier. This class uses binary split for both nominal and numeric attributes. For missing values, the method of fractional instances is used. This algorithm uses the both the gain index and gini index in calculating the best node in tree grown phase of the decision tree. This adds the best split node at the end of each phase. The best node is not the terminal nodes for splitting. It enables us to investigate new tree pruning methods that use cross-validation to select the number of expansions. Both pre-pruning and post- pruning is done in the same way.[10].

G. Spam Dataset

The spam dataset was taken from UCI machine learning repository and was created by Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt. Hewlett-Packard Labs. This dataset contains 4601 instances and 58 attributes (57 continuous input attribute and 1 nominal class label target attribute.[3].The class label has two values. 0- for not spam and 1-spam.

IV. EXPERIMENTAL RESULTS

Today, most of the data in the real world are incomplete containing aggregate, noisy and missing values. As the quality decision depends on quality mining which is based on quality data, pre-processing becomes a very important tasks to be done before performing any mining process. Major tasks in data pre-processing are performing feature reduction techniques. The feature reduction techniques used here are the ReliefF, ChiSquareAttributeeval, CFsubset evaluation methods. The Component Analysis is a dimension reduction technique which enables to visualize a dataset in a lower dimension without the loss of information.ReliefF algorithm detects conditional dependancies between attributes and provides a unified view on the attribute estimation in regression and classification. It is more robust and can deal with incomplete and noisy data.It evaluates the worth of an attribute by computing the value of chi-squared statistic with respect to class. The dataset is evaluated with 10-fold cross validations in the training data set. The various algorithms before filtering and after filtering is analyzed using the tables that are

©IJRASET 2015: All Rights are Reserved

Volume 3, Special Issue-1, May 2015
ISSN: 2321-9653

given below.

Table I: Details showing the performance of the classifiers in ReliefF filtering method

Algorithms	Test time (in sec)	Correctly classified instances (out of 4601 instances)	Accuracy (in %)	False positive (in %)
J48	0.28	4471	97.17	0.43
RndTree	0.3	4598	99.93	0
BFTree	0.22	4455	96.82	0.45
REPTree	0.25	4355	94.65	0.92
LMT	0.5	4534	98.54	0.34
SimpleCART	0.16	4431	96.30	0.67

Table II: Details showing the performance of the classifiers in CFsubsetEvaluation method

Algorithms	Test time (in sec)	Correctly classified instances (out of 4601 instances)	Accuracy (in %)	False positive (in %)
J48	0.26	4401	95.65	0.64
RndTree	0.28	4588	99.72	0
BFTree	0.16	4436	96.41	0.57
REPTree	0.35	4350	94.54	0.98
LMT	0.27	4336	94.24	0.99
SimpleCART	0.16	4322	93.94	1.02

Table III: Details showing the performance of the classifiers in ChiSquareAttributeeval method

Algorithms	Test time (in sec)	Correctly classified instances (out of 4601 instances)	Accuracy (in %)	False positive (in %)
J48	0.19	4477	97.30	0.41
RndTree	0.27	4598	99.93	0
BFTree	0.22	4451	96.74	0.45
REPTree	0.23	4357	94.69	0.94
LMT	0.92	4534	98.54	0.34
SimpleCART	0.16	4431	96.30	0.67

Table IV : Details showing the performance of the classifiers before using the filtering methods

Algorithms	Test time (in sec)	Correctly classified instances (out of 4601 instances)	Accuracy (in %)	False positive (in %)
J48	2.1	4278	92.98	1.56
RndTree	0.22	4184	90.93	2.25
BFTree	8.54	4267	92.74	1.42
REPTree	0.8	4274	92.89	1.48
LMT	771.35	4262	92.63	1.47
SimpleCART	8.33	4253	92.43	1.54

©IJRASET 2015: All Rights are Reserved

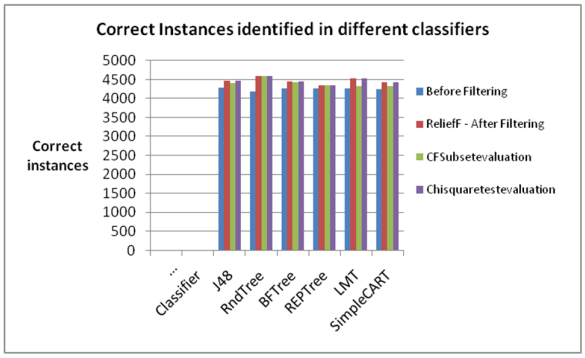


Fig.3 Results of correctly classified instances before and after using WEKA filters

A. Accuracy

Accuracy of a classifier was defined as the percentage of the dataset correctly classified by the method. This is given in Fig.4.

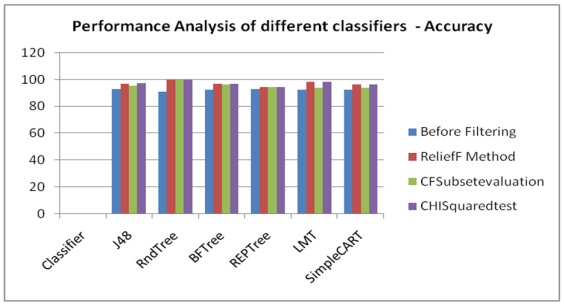


Fig 4 Results showing the accuracy of the classifier before and after using WEKA filters.

Accuracy is more than 96% (approx) in ReliefF and chi square evaluation method than in CF subset evaluation methods. The accuracy is 90% before filters are applied to the classifiers. The accuracy of the above algorithms are compared with each other before filtering and after filtering. ReliefF, CF subset evaluation method and Chi square attribute evaluation methods are used for spam filtering techniques for feature selection. Relief F filtering and Chi-square evaluation methods produce more correct instances than the CF Subset evaluation method. ReliefF filtering and Chi squared attribute evaluation yields best results for the two classifiers RndTree algorithm and LMT algorithm.

B. Error rate

Error rate of a classifier was defined as the percentage of the dataset incorrectly classified by the method. It is the probability of misclassification of a classifier

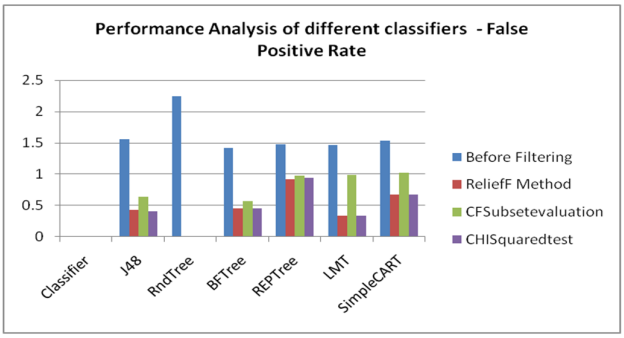


Fig 5 Results of the classifiers in predicting the False positive rate

The False Positive Rate for the above algorithms is specified and Rnd Tree algorithm showed the best in yielding 0% false positive rate. The accuracy of the above algorithm seems to be the best classifier among the other algorithms. The LMT algorithm showed 0.34% false positive rate in the ReliefF method and Chisquared test filters. The LMT algorithm has a great disadvantage with respect to the time taken to execute the test data set. The 10-fold cross-validation is applied to the data set but the LMT algorithm

works for hours together to produce the results.

C. Time

The Time taken for the algorithms to execute are tabulated and summarized in the fig 6.The LMT algorithm takes more time to execute than other algorithms. This algorithm has a great disadvantage in completing time.[12].

The 10-fold cross validations are done for each algorithm and LMT takes more than an hour to execute (771.35sec) in the training data set. The details of the time taken to execute the algorithms are stated in the Table V and the corresponding chart is given in Fig.6.

TABLE V: Details showing the execution time of the classifiers

Classifier	Before filtering	After Filtering – test data		
		Relieff	CF Subset	ChiSquare
J48	2.1	0.43	0.64	0.41
RndTree	0.22	0	0	0
BFTree	8.54	0.45	0.57	0.45
REPTree	0.8	0.92	0.98	0.94
LMT	771.35	0.34	0.99	0.34
SimpleCART	8.33	0.67	1.02	0.67

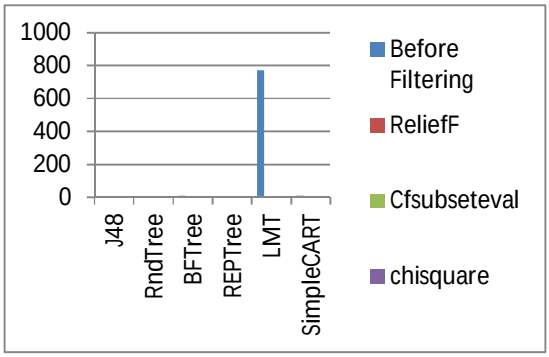


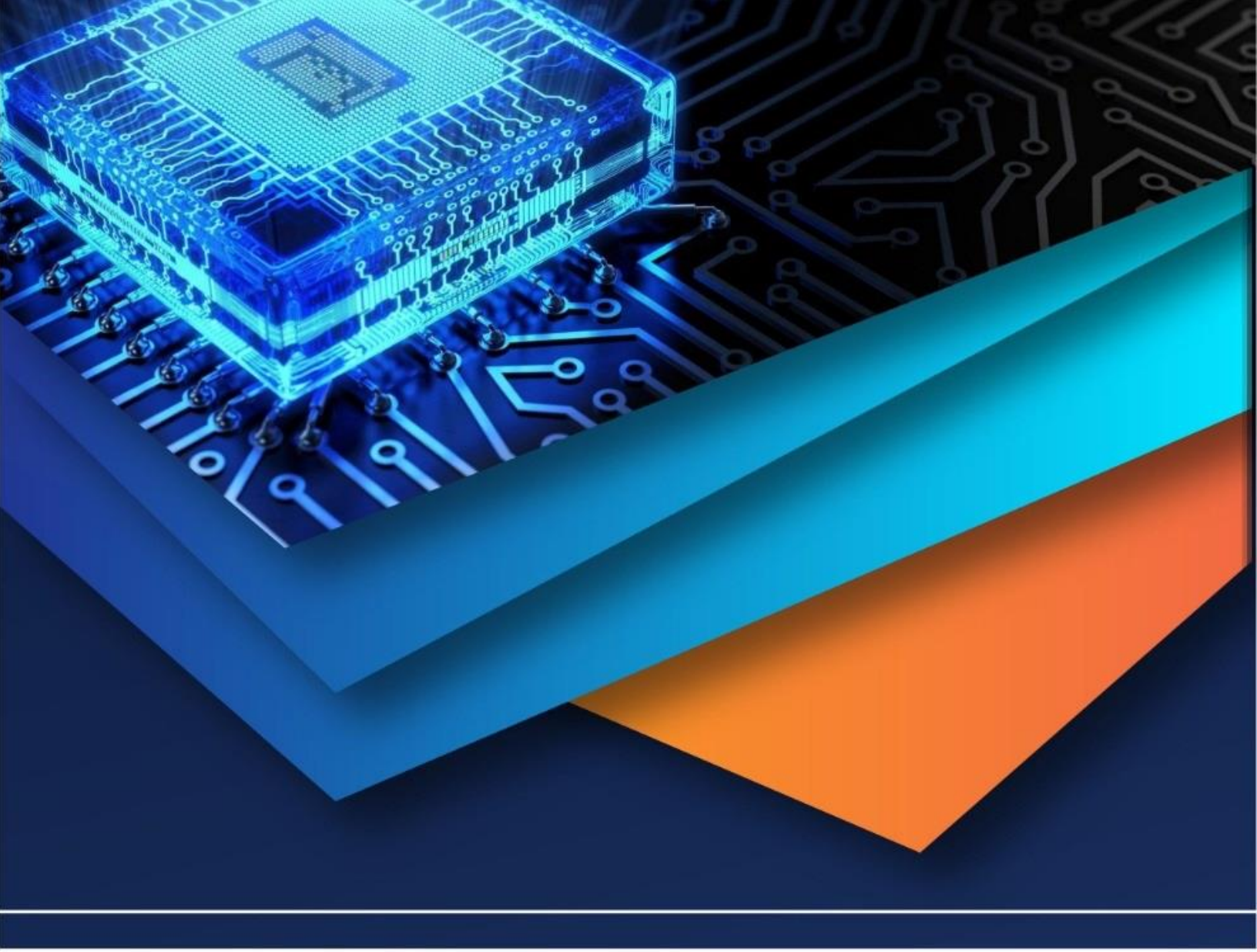
Fig.6 Results of the classifiers related to the execution time

V. CONCLUSION

E-mail spam classification needs more attention to identify the major threats and reduce the unwanted information from the spammers. Many researches have been going on to identify the best classifier in spam filtering. Among all the decision tree classifiers compared in this paper, the execution time, accuracy and low false positive rate has been exhibited only in Rndtree classifier. The accuracy of RndTree is 99% than the LMT classifier with an average of 98% and with the false positive of 0.34% in Chisquare and ReliefF filtering Techniques. The RndTree Classifier shows best performance than other decision tree classifiers.

REFERENCES

[1] Androutsopoulos, I.; Koutsias, J.; Chandrinos, K.Paliouras, G and Spyropoulos, C. 2000. An evaluation of naive bayesian anti-spam filtering.
[2] R.Kishore Kumar , G.Poonkuhali , P.Sudhakar, "Comparitive study on E-mail Spam Classifier Using data Mining Techniques" Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol 1. March 14-16, 2012 HongKong.
[3] UCI Machine Learning Repository – Spambase dataset http://archive.ics.uci.edu/ml/datasets/Spambase
[4] Patrick Ozer, "Data Mining Algorithms for classification,Radboud University Nijmegen, Jan 2008.
[5] Weka. WEKA (Data Mining Software)Available at: http://www.cs.waikato.ac.nz/ml/weka/.2006
[6] http://monkpublic.library.illinois.edu/monkmiddleware/public/analytics/clusterclassification.html
[7] http://searchmidmarketsecurity.techtarget.com/definition/spam-filter
[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] http://www.cs.waikato.ac.nz/~eibe/pubs/LMT.pdf
[10] http://www.uniroma2.it/didattica/WmIR/deposito/dectree_weka_tutorial.pdf
[11] Shomona Gracia Jacob, R.Geetha Ramani, Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.
[12] N. Landwehr, M. Hall, and E. Frank. Logistic model trees, 2003.
[13] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2005.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)