DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5571

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

CLASSIFICATION TECHNIQUES USING SPAM FILTERING EMAIL

Dr. C. Chandrasekar
Assistant Professsor
Department of Computer Science
Government Arts College, Udumalpet, India

P. Priyatharsini
M. Phil., scholar
P.G. and Research Department of Computer Science.,
Government Arts College, Udumalpet, India

Abstract: The general data mining model with the complex sample data solves the problem on data classification. The preprocessing step of complex data in data mining solves the problem of accuracy caused by the mass data.
The growing volume of spam mails annoys people and affects work efficiency significantly. The work focused on developing spam filtering algorithm, using statistics or data mining approach to develop precise spam rules. The main propose of an anti spam approach combining both data mining and statistical test approach. The efficiency of spam rules, only significant rules will be used to classify emails and the rest of rules can be eliminated for performance improvement.
The effective decision tree classifiers are used to classify whether the mail is spam or ham. Various 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. Two decision tree algorithms that are basically used as classifiers namely J48 or C4.5, Rndtree. The algorithms are 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 it shows best results among other classifiers.

Keywords: classifiers, e-mail, ham, spam

1. INTRODUCTION

Spam is an unfortunate problem on the internet.Spam emails accumulate in the users inbox without the consent of the user. Great issue the users inboxes are flooded up with spam mails and spend unproductive hours in deleting unwanted emails.Causes a loss of internet performance and wasting the network bandwidth, clogs up email servers to the point it sometimes crashes. Spam increases the spread of malware and viruses that pose a big threat to the network security and personal privacy. Spammers also deploy spam to gain personal information about the user for fraudulent purposes. The growing threats of spam definitely require drastic control measures.
Decision tree learning is a method for approximating discrete values target functions, in learned function is represented by a decision tree. [1]Decision tree learning is generally suited to the problems in that task is to classify into one of a discrete set of possible categories are often referred to as classification problems. A decision tree is used as a classifier for determining an appropriate action for a given case. Problem is process of determining whether the mail is Spam or Ham is to be found out. The mails that are identified as Spam is to be filtered and legitimate messages should be allowed to be stored in the users mail box. The information about the mail is given by a set of attributes such as frequent occurrence of the characters, words and special characters. The allowed actions are viewed as classes, can classify the mails as Spam or Ham.
A decision tree is a decision support tool that used for making decision analysis or changes of outcomes or to identify a strategy that is most likely to reach a goal. [7]. In particular, constructing classifiers in the form of decision trees has been quite popular, and a number of successful real world applications that employ decision trees construction have been reported.

What is a Spam Filter?

The task of Spam filtering is to rule out unsolicited mails automatically from a user's mail stream[2]. The various decision tree classifiers are taken for evaluation and apart from other types of data mining classifiers it is emphasized specifically on decision tree classifiers for the particular application of spam filtration technique. The main task of the spam filtration is to identify whether the mail is spam or not. [3]The decision tree filters are easy to implement and easy to understand. Provides an overall satisfactory performance as far as spam mail detection is concerned. The dataset is trained and tested with various decision trees and the performance evaluation criteria of various classifiers are based on the precision, accuracy and time taken by the classifier. The classifier which is evaluated best is further enhanced to provide more accuracy and the algorithm is implemented in the WEKA tool.

2. METHODOLOGIES

Muthukaruppan et al (2011) proposed method for hybrid scheme solves the problem of poor naive Bayes performance in a domain with dependent attributes, and the memory consumption problem of the decision tree. [13]The naive Bayes model at a leaf node should contain all the remaining attributes, large number of irrelevant attributes can be eliminated.
Kishore Kumar et al (2012) has taken spam dataset from UCI machine learning repository is taken as input data for analyzing the various classification techniques using TANAGRA data mining tool. [8]The various classification algorithms are applied over this dataset and cross validation is done for each of these classifiers.
Ruan Guangchen et al (2012) has used three types of decision tree classifiers such as Naive Bayes Tree [15]Classifier (NBT), C4.5, and Logistic Model Tree

Classifier were analyzed for Spam filtration. Among several approaches, the top most are SVM[12] (Support Vector Machines) and the well known Naive Bayes classifier. 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. Data set of emails is created using attributes and relations from the spam mails received in the mailbox for over six months. The 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 are run using Weka are NBTree, C4.5 decision tree classifier and Logistic Model Tree classifier are analyzed based on the performances with different criteria in terms of time, result efficiency and accuracy achieved by the various decision tree classifiers and also some other criteria like false positive, false negative rates of decisions taken by the classifiers.

Catarina Silva et al (2012) using hybrid system for text classification based on the ensemble of both Artificial Immune Systems (AIS) and SVM approaches. [6]The advantage of a non-evolutionary implementation that produced remarkable results with text classification and showing the classification performance gains, resulting in a classification has improved.

Manjusha et al (2013) used method for Binary Decision Tree Multi Class Support Vector Machine approach are using the advantages of SVM and decision tree[11], that is Decision Tree (DT) s are much faster than SVM s in classifying new instances while SVM perform better then DTs in terms of classification accuracy. To include both this advantages we will reduce the size of record set will be fed to the SVM. Normal data points are classified by decision tree while some crucial data points were difficult for decision tree to classify to multiclass SVM.

Malti Sarangal (2014) proposed method is K Means clustering and Support Vector Machine (SVM) based classification algorithm are considered to classify the spam base dataset[10]. The main advantages is improved classification accuracy and reduces the false positive and time cost. K Means algorithm, is numerical and one of the hard clustering method, this means that a data point can belong to only one cluster.

The decision tree classifiers provide great results as far as spam detection concerned. By comparing all the three classifiers, yield best results and provides 90% accuracy in performance. That algorithm takes more processing time than that of other classifiers. The one of the most disadvantages exhibited in this classifier. This is better than the earliest algorithms such as Naive Bayes and many other spam detection techniques.

3. EXISTING SYSTEM

The Immune System evolved to become an extremely complex resistance system that has the capability to identify foreign substances and to differentiate between harmless and harmful. Immune System is decomposed in two main layers of resistance that is innate and adaptive. Innate recognizes precised substances and its conduct is similar to all individuals of the same species. Adaptive layer is able to learn to identify new forms of anomalous pathogens that regularly change during the time hence it provides an extremely complicated adaptive form of identification.

The Immune System is also supported by a pathogens are divided into small peptides by Antigen Presenting Cell (APC). The peptides are then accessible by the lymphocytes also called as Transaction Cells. The Transcation cells have a particular set of receptors that used to bind peptides with a certain degree of affinity that are being offered by Antigen Presenting Cells. Artificial Immune Systems (AIS) is an adaptive system inspired by biological immune system and it is based on theoretical immunology.

4. K MEANS CLUSTERING

Automated mechanism uses unsupervised learning for classification purposed. Unsupervised learning means there is no supervisor is needed to train the mechanism. [4]Clustering is one type of unsupervised learning. Clustering is designed to aim for grouping similar type of data together. Clustering process data is divided into similar type of groups where each group contains the data which have more similarity. The groups are called as clusters. K Means clustering is the most useful method for finding natural groups of similar type of data.

A classification technique the objects are assigned to predefined categories whereas in clustering the classes are formed and two categories available for dividing clustering methods on the basis of character of the data and the reason for that cluster has being used. The categories are fuzzy clustering and hard clustering in the fuzzy clustering to every data element can belong to more than one cluster. Resolve it fuzzy clustering uses a mathematical model for classification and hard clustering every data element is divided into separate cluster.

K Means clustering algorithm is a hard clustering method so it can be applied for spam filtering. [14]The research utilized the K Means clustering algorithm to classify the emails. Classifies incoming email as spam or legitimate on the basis of similar attributes or features. The K Means clustering K is a positive number initialized in the starting and algorithm refer it to as the number of clusters required for classification. K Means clustering algorithm inspects the feature vector of each incoming email, such that the items within every cluster are similar to each other. The basis of this inspection it form two clusters, one is spam and another is legitimate. The iterative process where initial set of clusters and the clusters are frequently updated until no more upgrading is possible or the number of iterations reached to a specified limit.

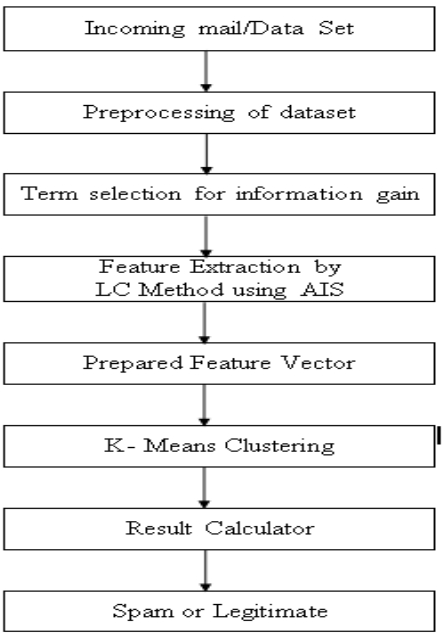


Figure 1.1 An overview of Local Concentration Based K Means Clustering

The local concentration based feature extraction method with artificial immune system has five processing stages are involved to generate final results. Each of them is discussed given blow.

Preprocessing of incoming email is essential task before process to classify it. The setup is working with real time spam filter, incoming email is processed and when working in an experimental environment sample datasets are preprocessed. Used string tokenizer in this phase for generating dictionary of the words. Irrelevant words are discarded and after it processed data is passed to term selection stage of the model.

Information Gain is used as term selection strategy for our model. [5]Algorithm for term selection is discussed as ds generation and term selection algorithm given below.

Step 1 : Initialize preselected set and DS == Empty set.

Step 2 : Every term in the terms set Do Calculate weight of the term according to a certain term selection strategy End

Step 3 : Arrange the terms in decreasing order of the weight

Step 4 : Join the front % terms to the preselected set

Step 5 : For all terms in the preselected set Do

Calculate Tendency as $(t_k)=P(t_k|c_l)$-P $(t_k|c_s)$

if $|| P(t_k|c_l)–P(t_k|c_s)||>$, >=0 then

if $|| P(t_k|c_l)$ –P $(t_k|c_s)$ || > , >=0 then

Add the term to DS_s

Else Add the term to DS_l

endif .

Else Discard the term

endif

endfor

P $(t_k|c_l)$ is probability of t_k as legitimate

P $(t_k|c_s)$ is probability of t_k as spam.

DS_s is spam detector set and

DS_l spam detector set.

Model used local concentration based feature extraction approach with artificial immune system. Algorithm used for feature extraction is discussed to local concentration based feature extraction approach with artificial immune system.

Step 1 : Move a sliding window of w_n term length over a given message

With a step of w_n term.

Step 2 : for every position of the sliding window Do Calculate the spam genes concentration in the window by formula: $SC_j = N_s/N_t$

Calculate the legitimate genes concentration of the window by formula: $LC_j = N_l/N_t$

end for.

Step 3 : Construct feature vector: ($<SC_1$, $LC_1>$,$<SC_2$,$LC_2>$…$<SC_n$, $C_n>$)

SC_j is spam gene concentration in j^{th} window.

LC_j is legitimate gene concentration in j^{th} window.

N_t is the number of dissimilar terms in the window.

N_s is the number of the dissimilar terms in the window which corresponding to detectors in D_s.

The work applied KMeans clustering for classification. Fourth and very important stage of spam filtering. The stage of measuring to effectiveness in this entire system by evaluating classification result. Algorithm used for K Means clustering at classification phase is discussed K Means clustering for classification

Step 1: Initialize spam and legitimate Centroids

Step 2: Centroids = kMeansInitCentroids(X, k)

Step 3: for iter = 1 iterations Cluster assignment step Assign each data

point to the closest centroid. idx(i) corresponds to cˆ(i), the indexof the centroid assigned to example i

Step 4: idx = findNearestCentroids(X, centroids); Move centroid step

Compute means based on centroidassignments

Step 5: centroids = computeMeans(X, idx, K)

Step 6: end

5. VARIOUS CLASSIFIERS IN EMAIL SPAM FILTERING PROBLEM DEFINITION

The various decision tree classifiers are taken for evaluation and apart from other types of data mining classifiers it is emphasized specifically on decision tree classifiers for the particular application of spam filtration technique. The main task of the spam filtration is to identify whether the mail is spam or not. The decision tree filters are easy to implement and easy to understand. Provides an overall satisfactory performance as far as spam mail detection is concerned. The dataset is trained and tested with various decision trees and the performance evaluation criteria of various classifiers are based on the precision, accuracy and time taken by the classifier. The classifier which is evaluated best is further enhanced to provide more accuracy and the algorithm is implemented in the WEKA tool.

6. SPAM DATASET

The spam dataset was taken from UCI machine learning repository and was created by Mark Hopkins et al Hewlett Packard Labs. The dataset contains 4601 instances and 58 attributes 57 continuous input attribute and 1 nominal class label target attribute. The class label has two values 0 for not spam and 1 for spam.

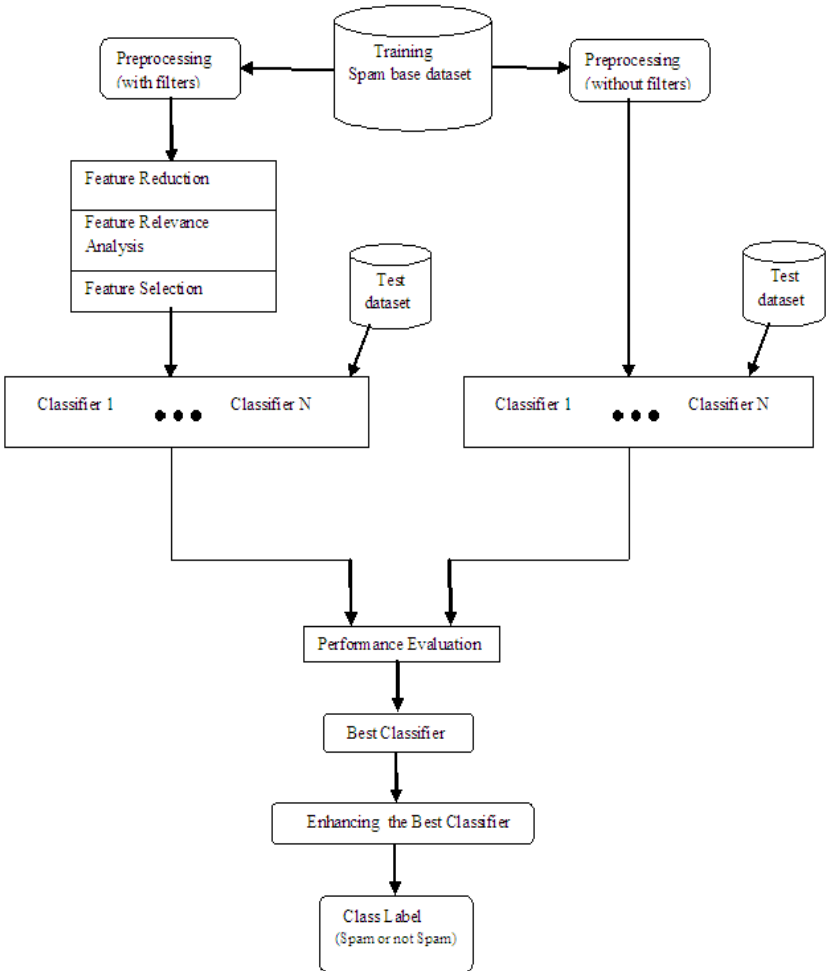


Figure 1.2 Various Classifiers in Email Spam Filtering

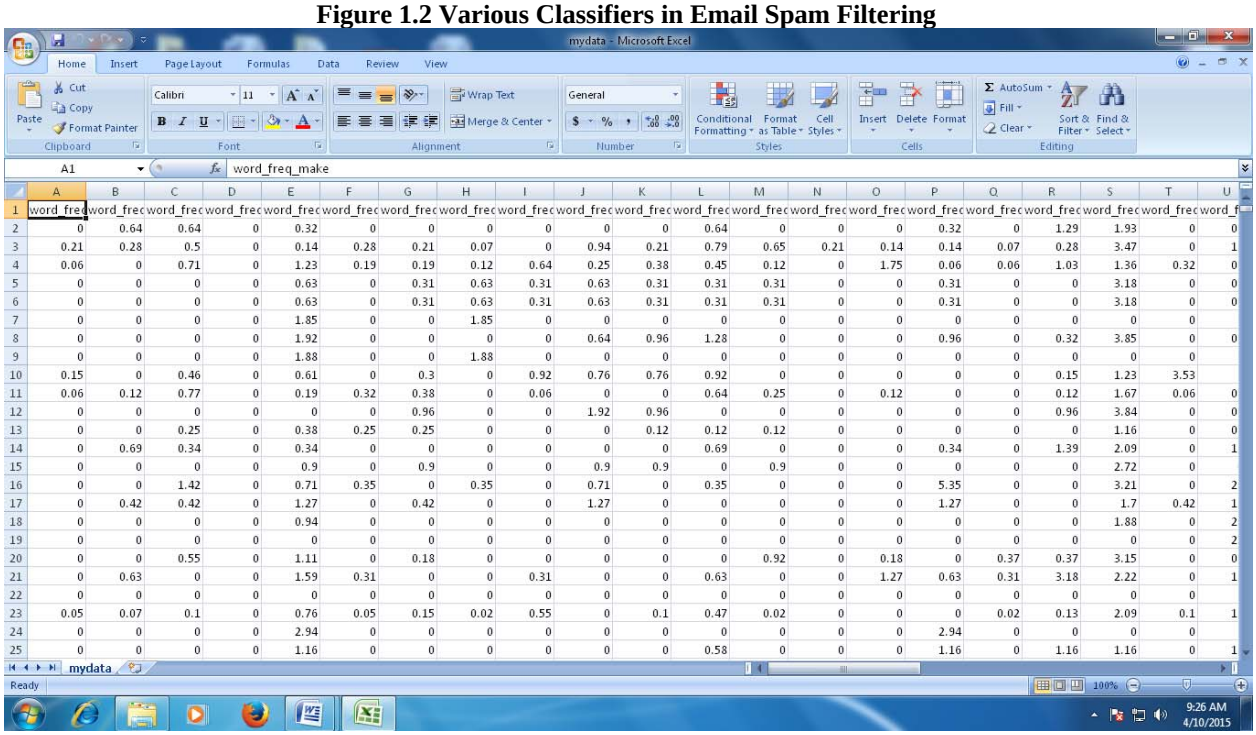


Figure 2.1 Spam Dataset

7. FEATURE REDUCTION TECHNIQUES

Complex data analysis and mining on huge amounts of data take a very long time making practical analysis infeasible. [12]Feature reduction techniques have been helpful in analyzing reduced representation of dataset without compromising the integrity of the original data and produce

quality knowledge. Feature reduction techniques reduce the volume of data or reduce the dimensions reduce attributes.

The feature reduction techniques used here are the ReliefF, ChiSquare Attribute evaluation, CFsubset evaluation methods. The Component Analysis is a dimension reduction technique re enables to visualize a dataset in a lower dimension without the loss of information.

ReliefF algorithm detects conditional dependencies between attributes and provides a unified view on the attribute estimation in regression and classification. The robust and can deal with incomplete and noisy data. Evaluates the worth of an attribute by computing the value of chi squared statistic with respect to class. The dataset is evaluated with ten fold cross validations in the training data set and tested.

Chi Square is a statistical test that measures the occurrence of features against the expected number of the occurrences of those features. The Chi Square evaluation method, the independent variables are the features and the dependent variables are the categories that is legitimate and spam email.

$$\chi^2 = \frac{N*(AD-CB)2}{(A+C)*(B+D)*(A+B)*(A+D)} \quad (4.1)$$

CFS Correlation based Feature selection Subset evaluation method uses a search algorithm along with a function to evaluate the merit of feature subsets. The heuristic by which Correlation based Feature selection Subset evaluation method measures the goodness of feature subsets takes into account the usefulness of individual features for predicting the class label along with the level of inter correlation among them. Correlation based Feature selection Subset evaluation method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

C4.5 ALGORITHM

Input: data training samples; list of attributes;
attribute_selection_method.
Output: decision tree.
Method:
Step 1: create a node N,
Step 2: if samples has the same class, C, then
Step 3: return N as leaf node with class C label;
Step 4: if list of attributes is empty then
Step 5: return N as leaf node with class label that is the most class in the samples.
Step 6: Choose test-attribute, that has the most Gain Ratio using attribute_selection_method;
Step 7: give node N with test-attribute label;
Step 8: for each ai pada test-attribute;
Step 9: Add branch in node N to test-attribute = ai;
Step 10: Make partition sample si from samples where test-attribute= ai;
Step 11: if si is empty then
Step 12: attach leaf node with the most class in samples;
Step 13: else attach node that generate by Generate_decision_tree si,attribute-list, test-attribute;
Step 14: endfor
Step 15: return N;

8. RANDOM FOREST ALGORITHM

Random Forest are ensemble of un pruned binary decision trees, unlike other decision tree classifiers Random Forest grows multiple trees are creates a forest like classification. Algorithm can be used for classification and regression.

Steps in Random Forest Algorithm:

Step1: A random seed is chosen which pulls out at random a collection of samples from training data set while maintaining the class distribution.

Step2: Selected dataset, a random set of attributes from the original data set is chosen based on user defined values.

Step3: A dataset M is the total number of input attributes in the dataset, only R attributes are chosen at random for each tree R<M.

Step5: The attributes from this set creates the test possible split using the Gini index to develop a decision tree model.

Step6: Random Forest Tree follows the same methodology and constructs multiple trees for the forest using different set of attributes.

9. EXPRIMENTAL RESULT

The discussions made in the project , they have created a dataset with different rules for finding whether the mail is spam or ham and they are implemented using 2 different classifier and J48 classifier. The results of the classifier are compared with the two different datasets and proved that J48 outperforms all the classifiers with 86% of accuracy and low false positive rate. The time taken by the J48 classifier is comparatively less than that of NB tree classifier.The results are compared and RndTree algorithm shows almost 99% accuracy level in filtering the spam mails and it shows best results among other classifiers.

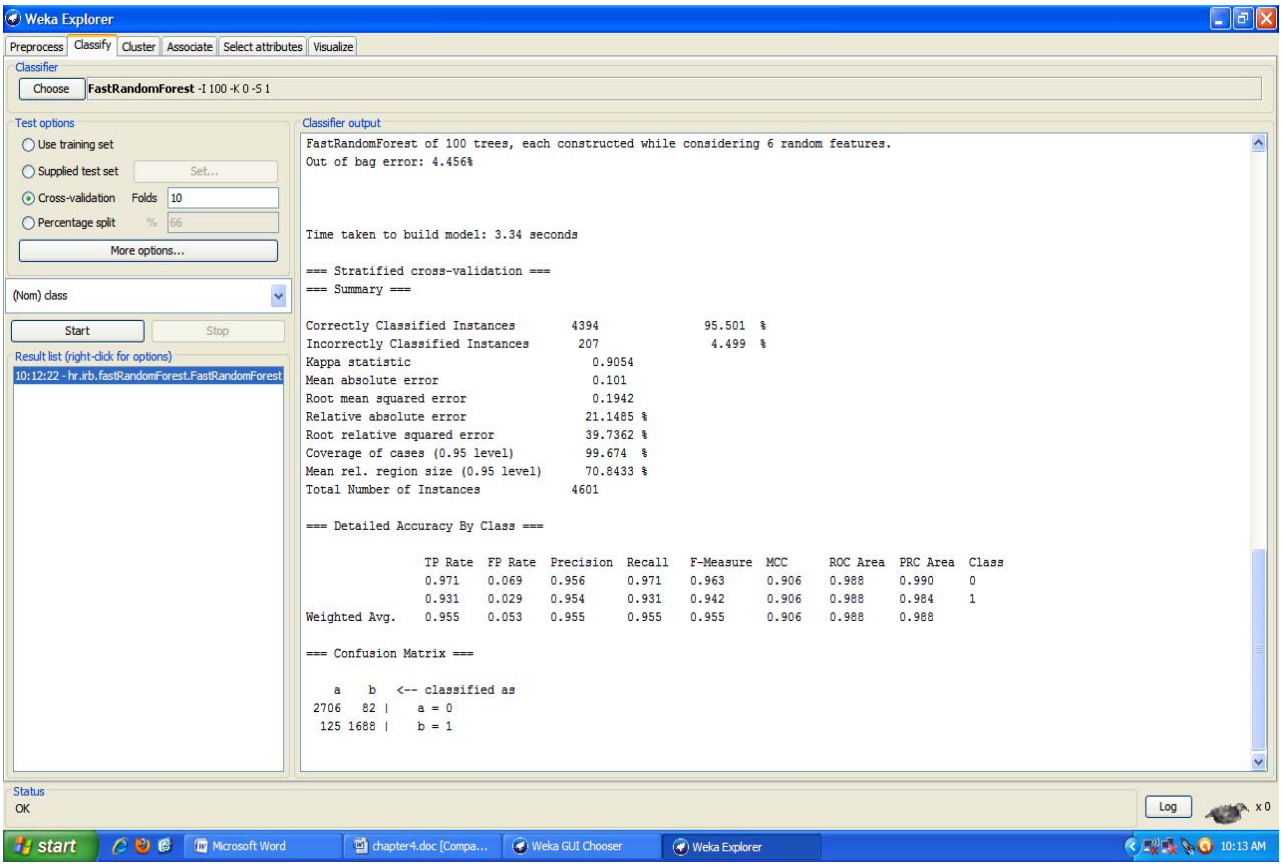
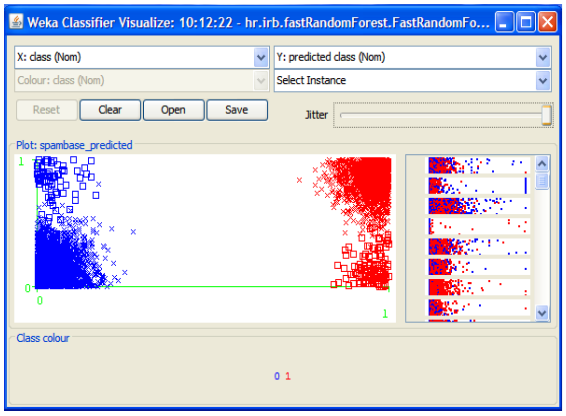
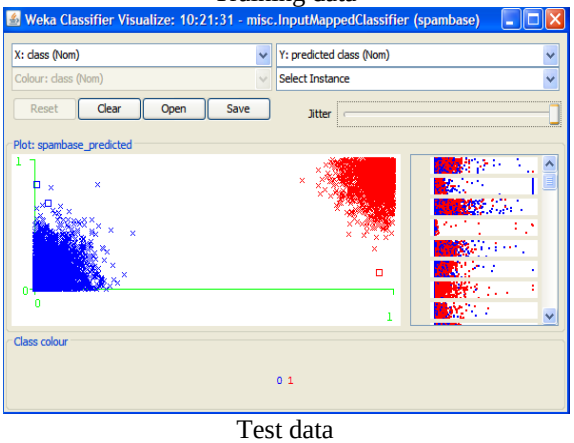


Fig 2.2: Enhancing Random forest Classifier



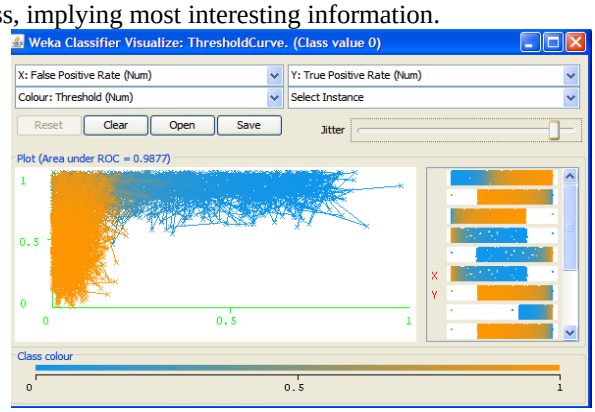
Training data



Test data

Figure 2.3 Calculate the Classifier Errors

Maximum 1 - 1/nc when records are equally distributed among all classes, implying least interesting information ‹ Minimum 0.0 when all records belong to one class, implying most interesting information.



Training data

Test data

Figure 2.3 Receiver Operating Characteristic Curve

ROC graphs are two dimensional graphs in which TP rate is plotted on the Y axis and FP rate is plotted on the X axis. An ROC graph depicts relative trade offs between benefits true

positives and costs false positives. Figure 2.3 shows an ROC graph with five classifiers labeled are all attributes.
The true positive rate and false positive rate are calculated by using logistic model cross validation method. Precision and recall are calculated for true positive and false positive rate.

Classified result is obtained by resulting input data and accurate result is formed by Logistic Model tree and construct the decision tree.

Table 1.1:Random Forest Classifier

Algorithms	Training data			Test data		
	Accuracy	Error rate	Precision	Accuracy	Error rate	Precision
Random forest	94.41	5.58	0.944	99.60	0.03	0.996
Enhanced Random forest	95.50	4.49	0.955	99.93	0.06	0.999

The results from the random forest classifier and enhanced random forest classifier for email spam filtering are tabulated in the table above. The results are taken before applying WEKA filters and the dataset is trained and tested for identifying the spam mails. The enhanced random forest classifier produces about 1.09% increased accuracy from the random forest classifier.

The error rate of random forest classifier is reduced to 4.49% in enhanced random forest classifier. While the data is tested, the accuracy of the classifier is further improved to 99.93% that achieves it as the best spam filtering algorithm. Algorithm is proved that the enhanced random forest classifier shows best results for spam filtering. The enhanced random forest classifier shows best precision rate of 0.999% while the dataset is tested.

The time taken by the various classifiers before and after applying WEKA filters in training the dataset is given in the table below.

Table 1.2 WEKA filters result for time

Algorithms	Before Filtering	After Filtering using WEKA filters		
		CFSubseteval	ReliefF Filtering	Chisquared attribute eval
	Training time (in Sec)	Training Time (in Sec)	Training time (in Sec)	Training time (in Sec)
C4.5/J48	0.84	0.31	0.89	0.72
Randomforest	0.74	0.44	0.55	0.53
Naivebayes	0.11	4.91	83.16	67.95

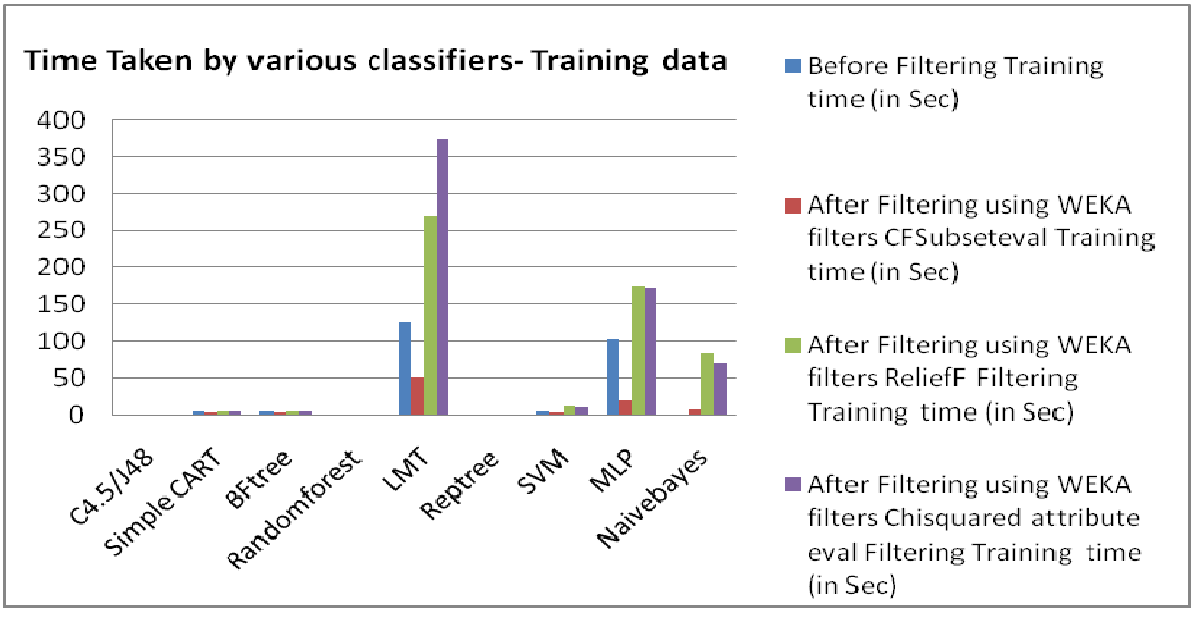


Figure 3.1 Results time taken by the classifier during the training time

Table 1.3 Naive Bayes classifier passed through the WEKA filters.

Algorithms	Before Filtering	After Filtering using WEKA filters		
		CFSubseteval	Relief Filtering	Chisquared attribute eval
	Test time (in Sec)	Test time (in Sec)	Test time (in Sec)	Test time (in Sec)
C4.5/J48	0.29	0.06	0.11	0.09
Randomforest	0.31	0.06	0.13	0.14
Naivebayes	0.20	0.09	0.20	0.20

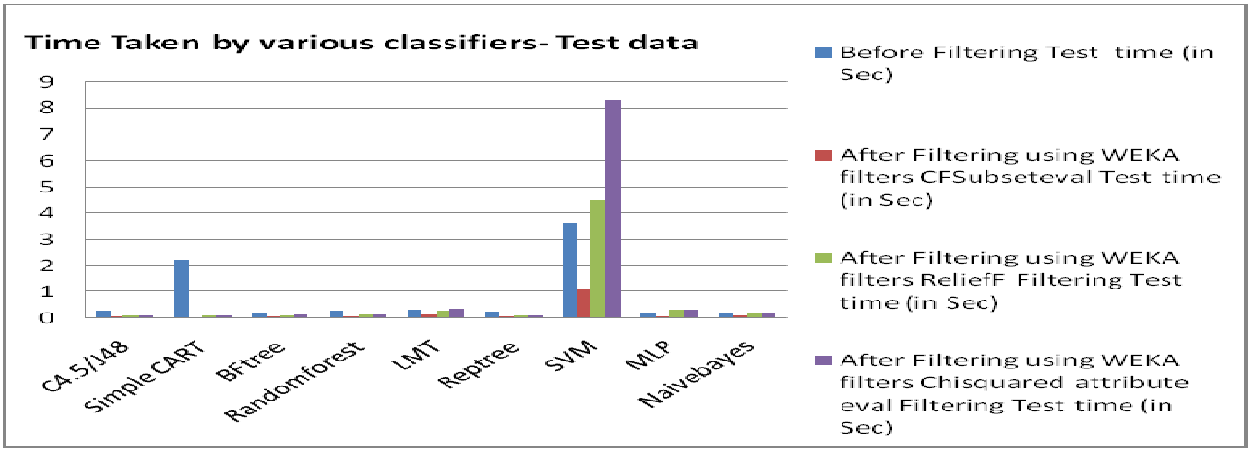


Figure 3.2 Time taken to test the dataset applying WEKA filters

Accuracy of the classifier is defined as the degree of closeness of classifying the correctly classified instances.

Table 1.4 Accuracy of the various classifiers

Algorithms	Before Filtering (in %)	After Filtering using WEKA filters (in %)		
		CFSubseteval	ReliefF Filtering	Chisquared attribute eval
	Training data	Training data	Training data	Training Data
C4.5/J48	92.97	92.69	93.00	92.97
Randomforest	94.41	93.71	94.47	94.41
Naivebayes	79.28	93.30	93.08	93.19

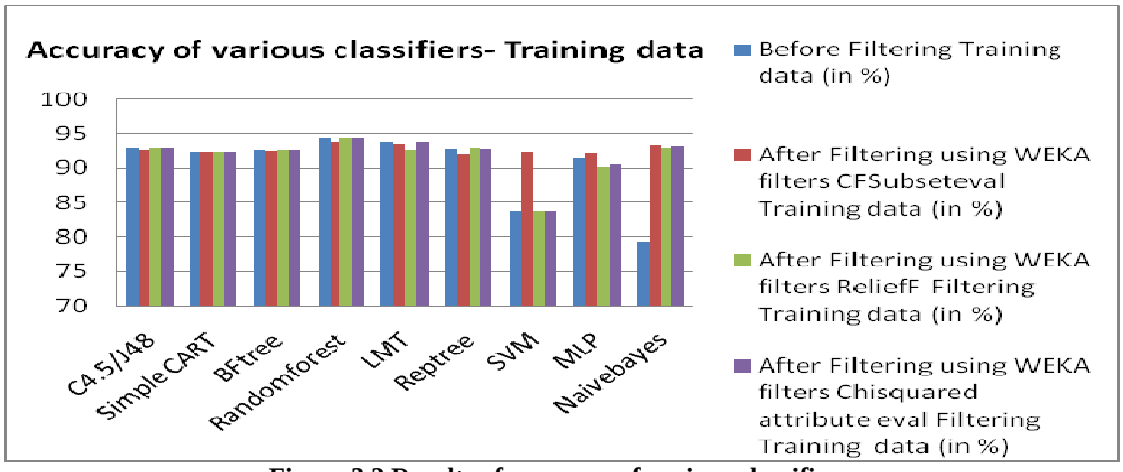


Figure 3.2 Results of accuracy of various classifiers

10. CONCLUSION AND FUTURE ENHANCEMENT

Email spam is a serious threat in corporate world and also in business. Reducing the spam mails and preventing the accumulation of spam mails storing in user's mailbox is a great challenge to the users. The identification of best algorithm to classify the spam mails is an important task.

Decision tree algorithms are used in filtering the spam mails because the main task is to classify the mails whether it belongs to spam or ham. The algorithms are trained, tested before and applying filtering algorithms. The results of the different algorithms are evaluated based on the Accuracy, Error rate, Precision and False positive rate. The comparison of the above algorithms based on their performance shows that the Random forest classifier exhibit best results when compared to other classifiers before and after applying weka filters.

The bugs that are identified when this classification algorithm was built are when handling with the missing values. Split point is the point at the tree splits up the instances in two instances by assigning weights to the branch at the splitting point. The attribute has some missing values, the attributes carry some information after the split points. The results in additional branches in the tree. Sometimes, the split will have a reduction of entropy of 0 and have a small positive value which leads to additional branches in the tree. The algorithm can be further enhanced by improving the Out of Bag estimate (OOB) it supports multithreading.

REFERENCES

[1] Abdelrahim, Adoutspouls, Agarwal, Awan and Roth (2000), 'Effect attribute size of training spam lists for Naive Bayesian Filters', IEEE Transaction Pattern Anal. Machine learning Intelligence., vol. 26, no. 11, pp. 1475–1490

[2] Ali Ahmed Abdelrahim, Ammar Ahmed, El hadi, Hamza Ibrahim (2000), 'Feature Selection and Similarity Coefficient Based Method Email Spam Filtering', International Conference on Computing, Electrical and Electronic Engineering., vol. 16, no. 21, pp. 245–250

[3] Amandeep Kaur and Malti Sarangal (2009), 'A Hybrid Approach For Enhancing The Capability of Spam Filter', International Journal of Computer ApplicationsTechnology and Research, Volume 2– Issue 6, ISSN 759 – 762

[4] Bay, Ess, Matsumoto, Tuytelaars and Gool (2004), 'Study on two spam detection metods for support vector machines and Naïve Bayesian classifier', Journal of Computing Machine learning algorithms ., vol. 110, no. 3, pp.346-349

[5] Caputo, Hayman and Xing Tan (2009), 'Colonel particle swarm optimization in multilayer neural networks and support vector machines ', in Proceedings International Conference on computing research techniques for learning algorithms ., vol. 2,pp. 1597–1604

[6] Catatrina Silva, Dollar, Wojek, Schiele, and Perona (2012), 'A Hybrid text classification based on Artificial immune system and support vector machine', International journal of computing and communication engineering, vol. 34, no.4 , pp. 743–761.

[7] Ichimura, Shechtman and Irani (2007), 'Spam classification in self organizing map and Automatically defined group', IEEE International. Conference on soft computing and fuzzy systems, pp 1-8

[8] Kishorkumar, Revathi and Chen (2012), 'A Comparative study of various data mining classification algorithms', IEEE Transaction on Intelligence computing, vol .32,no.32, no.9 , pp. 1705-1720

[9] Kunal Jain, Amrit Pal, Manish Sharma, Sanjay Agrawal (2014), 'Impact of Spam on the Environment & its Prevention', IEEE International Conference on Convergys of Technology, Pune., vol. 7, no. 6, pp. 567–575

[10] Malti Sarangal, Amandeep Kaur, RajeshKanna (2014), 'A KMeans Clusteringand Support Vector machine based classification algorithms of Spam Filter', International Journal of Computer Applications Technology and Research, Volume 12– Issue 31, ISSN 345 – 353

[11] Manjusha, Perona and Zisserman (2013), 'Email spam detection using Binary Decision Tree Multiclass Support Vector machines', in Proceeding of International. Conference computer vision and knowledge engineering., vol. 2, pp 264-271

[12] Mario Antunes, Catarina Silva, B ernardete Ribeiro, and Manuel Correia (2010), 'On Using An Ensemble Approach of AIS and SVM for Text Classification'., vol. 17, no. 9, pp. 98–104.

[13] Muthukaruppan Annamalai, Ankur Jain, Vaishn avi Sannidhanam (2011), 'A Novel Hybrid Approach to Machine Learning', Study Documents, Department of Computer Science and Engineering, University of Washington. vol. 2, no. 5, pp. 114-340

[14] Nadir Omer Fadl Elssied, Otman Ibrahim, Waheeb Abu Ulbeh (2014), 'An Improved of Spam Email Clas sification Mechanism Using KMeans Clustering', Journal of Theoretical and Applied Information Technology, Vol. 60 No.3

[15] Ruan Guangchen and Tan Ying (2012), 'A Three Layer Back-Propagation Neural Network for Detection Using Artificial Immune Concentration', SoftComputing,vol.14,pp.139–150