International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 12, December 2015

E-Mail Filtering For The Removal Of Misclassification Error

[1]Saurbh kumar [2] dr. Manish mann
L.r. Institute of engineering & technology, solan
Email: shaanushukla61@gmail.com

Abstract: — Email is electronic mail. It is method of exchanging digital messages from source to destination. The exchange of messages from an author to one or more. Email messages can be text files, graphics images and sound files. But now-a-days, the problem in the email is spam and security also. Text editor is included in the email systems to compose the messages. When one send the message to the on specified address then one can also send the same message to the several users and this is called broadcasting. Email filtering is the processing of email to systematize it according to the exact criteria. Most often this refers to the automatic processing of incoming messages, but the term is also used to the involvement of human intelligence in addition to anti-spam techniques. Bayesian spam filtering is a statistical method of e-mail filtering. Bayesian spam filtering makes use for Naive Bayes classifier to make out spam e-mail. Work is classified by Bayesian to compare the use of tokens i.e typically words, or we can say irregularly other things, with spam and non-spam e-mails. Bayesian spam filtering is a extremely powerful technique for constricting with spam, that can adapt itself to the email needs of individual users, and gives low false positive spam finding rates that are generally acceptable to users. Our purpose is to reduce the misclassification error.

Keywords:Anti-spam,Email filtering techniques, Naive Bayes, Misclassification Error, Spam.

I. INTRODUCTION

Email filtering is the processing of email to systematize it according to the exact criteria. Most often this refers to the automatic processing of incoming messages, but the term is also used to the involvement of human intelligence in addition to anti-spam techniques. Bayesian spam filtering is a statistical method of e-mail filtering. Bayesian spam filtering makes use for Naive Bayes classifier to make out spam e-mail. Work is classified by Bayesian to compare the use of tokens i.e typically words, or we can say irregularly other things, with spam and non-spam e-mails. Bayesian spam filtering is a extremely powerful technique for constricting with spam, that can adapt itself to the email needs of individual users, and gives low false positive spam finding rates that are generally acceptable to users.

Bayesian filtering is one of the most effective and bright solutions to fight with spam email nowadays. Spam is a trouble faced by all email users and it reflects no sign of slowing down anytime shortly; in fact, the number of spam emails is growing daily. Added to this, spammers are becoming more complicated and are continuously managing to outsmart 'static' methods of fighting spam.

There are basically two types of filtering, inbound filtering and outbound filtering. In inbound filtering, email messages are sheltered by the filtering system. In this type of filtering, message scanning process is involved. In case of outbound filtering, scanning email messages from local users before any potentially unsafe messages can be delivered to others on the Internet. Outbound email filtering is commonly used by Internet service providers is transparent SMTP, in which email traffic is intercepted and filtered by means of a transparent proxy within the network.

II.EMAIL FILTERING TECHNIQUES:

The techniques currently used by most anti-spam software are static, meaning that spammers.Simply examine the latest anti-spam filtering techniques and hit upon ways how to cut them, usually done by simply change the message a little. This gave anti spam developers a new challenge – come up with a new anti methods. spam technique; one that was familiar with spammers' tactics as they vary over time, and that is capable to adapt to the particular organization that it is protecting from spam. There are different emails filtering

1) Blacklist:

Blacklist comes under the list based filters. This is spam filtering method attempts to stop unwanted email by blocking messages from the list of sender. Blacklist contains the records of email addresses. In this when in coming message arrives, the spam filter checks to see if its IP or email address is on the blacklist. Then it considers the message as a spam and then reject it.

2) Whitelist:

Whitelist blocks spam using a system almost exactly opposite to that of blacklist. In this if an unknown sender's email address is checked against the database, if they have no history of spamming, their message is sent to inbox and then they added to the whitelist.

3) Word based filtering :

Word based filtering comes under the content based filtering it is the simplest form of filtering .word based filtering is the capable technique for fighting junk email. for example, if the filter has been set to stop all messages containing the word "acbd". But spammers often purposefully misspell keywords in order to evade word based filtering and this is the main problem in this type of filtering.

4) Heuristic filters:

This type of filtering contain multiple terms instead of containing one term based on the word based filtering. In this filter adds up all the points and then calculates the total score, the heuristic filters work fast.

5) Bayesian filters:

Bayesian filters technique is the most advance content based technique. It employs the laws of mathematical probability to settle on which message are real and which message is spam. In this, filter takes words and phrases finding legitimate mails ad adds them to the list. This method acquires a training time period before it starts running well.

B.Classification:

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Data classification is a two steps process, consisting of a learning step(where the classification model is constructed) and the classification step(where the model is used to predict class labels for given data).

C.Classification Algorithms:

Data Mining provides the following algorithms for classification:

1.Decision Tree:

A decision tree is a flowchart like tree structure, where each internal node (non-leaf node) denotes a test on a attribute, each branch represents an outcome of the test, and each leaf node(or terminal node) holds a class label. The topmost node in a tree is root node. A typical decision tree represents the concept buys computer , that is, it predicts whether a customer at All Electronics likely to purchase a computer. Internal nodes are denoted by rectangle, and leaf nodes are denoted by ovals. Some decision tree algorithm produce only binary trees (where each internal node branches to exactly two other nodes) , where others can produce non binary trees.

The Decision Tree algorithm produces accurate and interpretable models with relatively little user intervention. The algorithm can be used for both binary and multiclass classification problems. The algorithm is fast for both at build time and applies time. The build process for Decision Tree is parallelized. (Scoring can be parallelized irrespective of the algorithm.) Decision tree scoring is especially fast.

*2.Naive Bayes :*The Naive Bayes algorithm is based on conditional probabilities.It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and Arepresents the prior event.Naive Bayes handles missing values naturally as missing at random. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors. Missing values in nested columns are interpreted as sparse. Missing values in columns with simple data types are interpreted as missing at random. The Naive Bayes algorithm affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows. The build process for Naive Bayes is parallelized. (Scoring can be parallelized irrespective of the algorithm.)Naive Bayes can be used for both binary and multiclass classification problems.

*3.Support Vector Machine :*It is a method of classification of both liner and non-liner data. In nutshell, an SVM is an algorithm that works as follows. It uses nonlinear mapping to transform the orginal training data into higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (i.e., a "decision boundary" separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors("essention" training tupples) and margins(defined by the support vectors.

A standout amongst the most famous techniques or structures utilized by information researchers at the Rose Data Science Professional Practice Group is Random Forests. The Random Forests calculation is one of the best among order calculations - ready to arrange a lot of information with exactness. Irregular Forests are an outfit learning technique (likewise considered as a type of closest neighbor indicator) for characterization and relapse that build various choice trees at preparing time and yielding the class that is the method of the classes yield by individual trees (Random Forests is a trademark of Leo Breiman and Adele Cutler for a troupe of choicetrees). Arbitrary Forests are a mix of tree indicators where every tree relies on upon the estimations of an

irregular vector examined autonomously with the same circulation for all trees in the woods. The essential rule is that a gathering of "frail learners" can meet up to shape an "in number learner". Irregular Forests are a grand instrument for making expectations considering they don't overfit as a result of the law of extensive numbers. Presenting the right sort of haphazardness makes them exact classifiers and regressors. Single choice trees frequently have high change or high inclination. Irregular Forests endeavors to moderate the issues of high fluctuation and high inclination by averaging to locate a characteristic harmony between the two extremes. Considering that Random Forests have couple of parameters to tune and can be utilized basically with default parameter settings, they are a basic device to use without having a model or to create a sensible model quick and proficiently.

Arbitrary Forests are anything but difficult to learn and utilization for both experts and laypeople - with little research and programming obliged and may be utilized by people without an in number factual foundation. Basically, you can securely make more precise forecasts without most essential missteps basic to different strategies.

The Random Forests calculation was created by Leo Breiman and Adele Cutler. Arbitrary Forests develops numerous characterization trees. Every tree is developed as takes after:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

The reaction of every tree relies on upon an arrangement of indicator qualities picked freely (with substitution) and with the same conveyance for all trees in the woods, which is a subset of the indicator estimations of the first information set. The ideal size of the subset of indicator variables is given by log2 M+1, where M is the quantity of inputs. For grouping issues, given an arrangement of straightforward trees and an arrangement of arbitrary indicator variables, the Random Forest technique characterizes an edge work that measures the degree to which the normal number of votes in favor of the right class surpasses the normal vote in favor of some other class display in the reliant variable. This measure furnishes us not just with a helpful method for making expectations, additionally with a method for partner a certainty measure with those forecasts.

For relapse issues, Random Forests are framed by developing straightforward trees, each fit for creating a numerical reaction esteem. Here, as well, the indicator set is haphazardly chosen from the same appropriation and for all trees. Given the over, the mean-square blunder for a Random Forest is given by:

mean error = (observed - tree response)^2
The expectations of the Random Forest are taken to be the normal of the forecasts of the trees:
The expectations of the Random Forest are taken to be the normal of the forecasts of the trees:
The predictions of the Random Forest are taken to be the average of the predictions of the trees:

$$\text{Random Forest Prediction } s = \frac{1}{K}\sum_{K=1}^{K} K^{th} \text{ tree response}$$

Where the record k keeps running over the individual trees in the woods. Commonly, Random Forests can adaptably fuse missing information in the indicator variables. At the point when missing information are experienced for a specific perception (case) amid model building, the expectation put forth for that defense is in view of the last going before (non-terminal) hub in the particular tree. Thus, for instance, if at a specific point in the succession of trees an indicator variable is chosen at the root (or other non-terminal) hub for which a few cases have no legitimate information, then the expectation for those cases is essentially taking into account the general mean at the root (or other non-terminal) hub. Subsequently, there is no compelling reason to dispense with cases from the investigation on the off chance that they have missing information for a portion of the indicators, nor is it important to process surrogate part insights.

III. RELATED WORK

Most of the work is taking place in email filtering. Mostly the filtering for spams and hams in those algorithms are based on ‘content based’. Some of them are as follows:
SahamiMehran ,Dumais Susan et al (1998) In addressing the growing problem of junk Email on the internet a method is examined for the automated construction of filters to eliminate such unwanted messages from user’s mail stream. By casting this problem in decision theoretic framework it can be able to make used of probabilistic learning methods in conjunction with a notion of differential misclassification cost to produce filters which are especially appropriate for the nuances of this task.

while this may appear at first to be straight forward text classification problem, much more accurate filters can be produced to show that by considering domain specific features of this problem in addition to the raw text of email messages, finally it can be shown that all efficiency of such filters in a real world using scenario arguing that this technology is mature enough for deployment[1].

*KonstantinosV.Chandrinos.ConstantineD.Spyropoulos (2000)*In the proposed research, to detect the spam Naive Bayesian is trained automatically. This approach is tested on a large collection of personal email messages which are made publically available in "encrypted" from contributing towards standard benchmarks. Appropriate Cost sensitive measures are introduced. In this approach Naive Bayesian filter is compared to see the performance, to filter which is part of widely used email reader. In this approach filtering/routing, text categorization, test collection keywords are used. In conclusion, it concluded after experiment results that cost sensitive evaluation suggests that neither the Naive Bayesian nor the keyword-based filter perform well enough to be used.

M. Basavaraju,Dr. R. Prabhakar et al(2012) A novel method of efficient spam mail classification using clustering techniques is presented in this research. E-mail spam is one of the main problems of the today's internet, bringing financial harm to companies and annoying individual users. In between the approaches developed to discontinue spam, filtering is an important and popular one. A new spam finding technique using the text clustering based on vector space model is proposed in this research paper. By using this method, one can take out spam/non-spam email and detect the spam email efficiently. Vector space model shows the representation of data. Clustering is the technique used for data reduction. It splits the data into the groups based on pattern similarities such that each group is abstracted by one or more representatives [8].

3. Proposed Work Our research is for the less error prone classification by reducing the misclassification. Misclassification is defined as when legitimate emails are categorized as junk emails or vice versa. Cost of misclassifying legitimate emails as junk is much higher than the cost of junk mails as legitimate mails. Remedies can be found using the following steps:

- Classification scheme which will provide probability for itsclassification decisions
- Cost of these two kind of misclassification errors

The above concepts are implemented in the following algorithms for classification. These algorithms are:

- Naïve Bayes Classifier
- Decision tree

In case of Linear Discriminant Analysis, there are training data and sample data. The observations with known class labels are known as training data. There are sample data on which we will be using the training data sets. Then we will be computing the resubstitution error which is the misclassification error (the proportion of misclassified observations) on the training set. We will also compute the confusion matrix on the training set. A confusion matrix contains information about known class labels and predicted class labels. Generally speaking, the (i,j) element in the confusion matrix is the number of samples whose known class label is class i and whose predicted class is j. The diagonal elements which would be represented in graph will correctly classified observations.For some data sets, the regions for the various classes are not well separated by lines. When that is the case, linear discriminant analysis is not appropriate. Instead, you can try quadratic discriminant analysis (QDA) for our data. Decision trees can handle both categorical and numerical data. For the decision tree algorithm, the cross-validation error estimate is significantly larger than the resubstitution error. This shows that the generated tree over fits the training set. In other words, this is a tree that classifies the original training set well, but the structure of the tree is sensitive to this particular training set so that its performance on new data is likely to degrade. It is often possible to find a simpler tree that performs better than a more complex tree on new data.

The objective of our work is to minimize the classification error by reducing misclassification. As the base of our research is Naïve Bay's algorithm, so we will be implementing the Naïve Bay's algorithm at first. Our proposed method is based on decision tree, so we will be implementing the standard decision tree algorithm. The next phase is our modified decision tree algorithm. Implementation of our modified decision tree algorithm will be followed by the error detection of these three algorithms and the algorithm with least error will be chosen as the best way to filter emails. The steps are:

- Accessing and categorising the UCI repository on email filtering
- Implement Naive Bay's Algorithm
- Implement decision tree algorithm
- Finding out the misclassification error

IV.NEED OF OUR METHODOLOGY

Previously, spam classification is done on different classification algorithm and it was found that Random Forest algorithm is best suitable for the same. But there are some disadvantages of Random Forest algorithm. These are:

1. Large number of trees may make the algorithm slow for real-time prediction.
2. It is not suitable for less number of dataset due to longer execution time.
3. Hard to understand

As our dataset is already filtered, we will not need to create large number of trees. So from the angle of dataset, decision tree best suits our research. It has the following advantages:

1. Easy to interpret and explain
2. Lesser execution time over random forest

Methodology

Part A

1. port the dataset
2. Separately store the numeric data and the class labels
3. Scatter the data in the axis
4. Assign colours to spam and non-spam data
5. Classify the data
6. Calculate bad sector
7. Calculate linear re-substitution error
8. Plot the classified dataset
9. Calculate quadratic substitution error

Part B

1. Import the dataset
2. Separately store the numeric data and class labels
3. Scatter the data in the axis
4. Assign colour to spam and non-spam class labels
5. Calculate the Gaussian prediction
6. Calculate the bad sectors
7. Calculate the Gaussian re-substitution error

Part C

1. Import the dataset
2. Separately store numeric data and the class label
3. Scatter the data in the axis
4. Assign colours to spam and non-spam
5. Calculate cross validation.
6. Calculate 'naïveBayes Kernel Density re-substitution error.
7. Calculate Naïve Bayes Kernel Density cross validation error.
8. Plot the dataset.

Part D

1. Import the dataset
2. Separately store numeric data and the class label.
3. Scatter the data in the axis.
4. Assign colours to spam and non-spam.
5. Partition the dataset using cross validation.
6. Create a tree.
7. Find out the bad sector.
8. Calculate the decision tree re-substitution error.
9. Calculate the decision tree cross validation error
10. Calculate the best level
11. Prune the tree to enhance its efficiency
12. Calculate the cost

Part E

Find the best suitable algorithm for classification of spam data depending upon the errors of each algorithm calculated above.

The dataset for the implementation is loaded. The numeric data is imported to the dataset variable and the class labels are stored in mailgroup variable.The dataset is taken from the UCI repository dated 1 July 1999

Dataset-Characteristics	Multivariate	**Number of Instances**	4601	**Area**	Computer
Attribute Characteristics	Integer, Real	**Number of Attributes**	57	**Date Donated**	1999-07-01

Dataset information of spam mails from UCI repository

The last column denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. Here are the definitions of the attributes:

48 continuous real [0,100] attributes of type word_freq_WORD
= percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type char_freq_CHAR]
= percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type capital_run_length_average
= average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_longest
= length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_total
= sum of length of uninterrupted sequences of capital letters
= total number of capital letters in the e-mail
1 nominal {0,1} class attribute of type spam

= denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

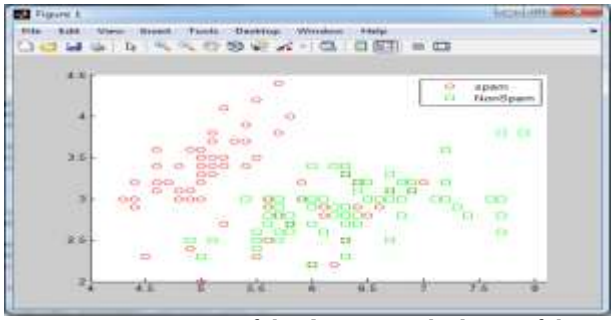

Figure 3.2 Scattering of the dataset on the basis of the class labels spam and Non-spam.

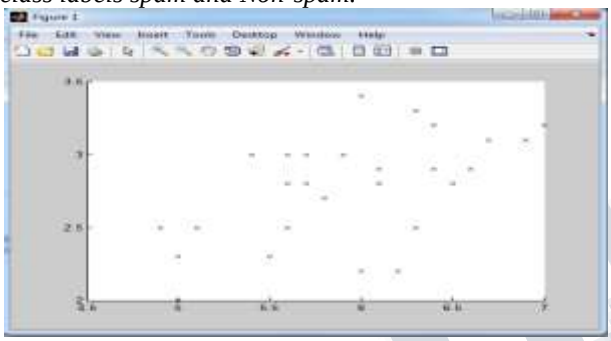

Figure 3.3 Misclassification plotted of Spam and

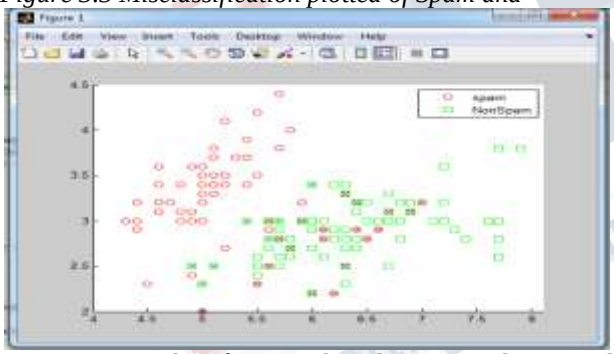

Figure 3.4 Misclassification plotted on original scattered class labels.

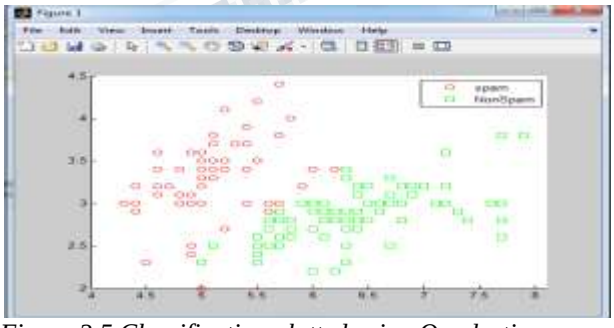

Figure 3.5 Classification plotted using Quadratic Distribution.

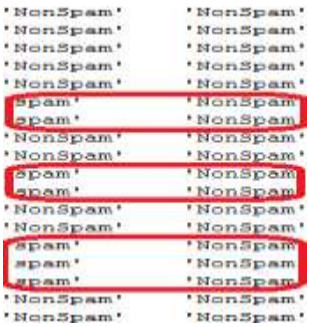

Figure 3.6 Display of Naive Bayes Gaussian distribution aganist original class labels.

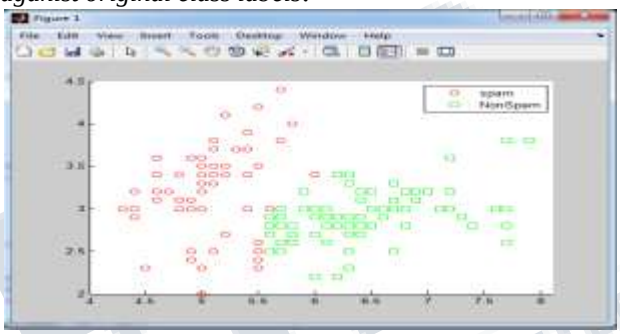

Figure 3.7 Classification using Naive Bayes Gaussian distribution.

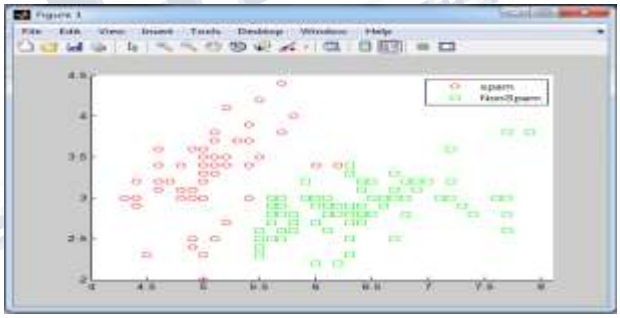

Figure 3.8 Classification plotted using Naive Bayes Kernel distribution.

Using Decision Tree classification technique a mesh grid was created first to define the border for the classification.

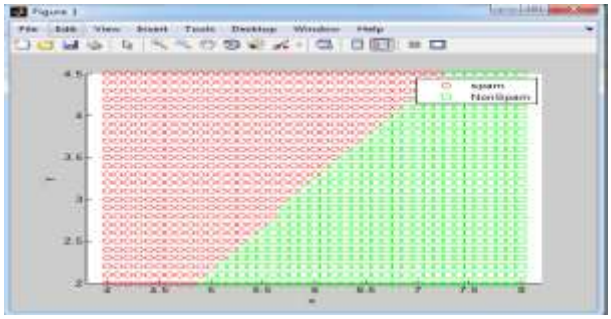

Figure 3.9 Scattering of mesh grid for x and y axis.

Classification begins with the initialization of dataset under decision tree.

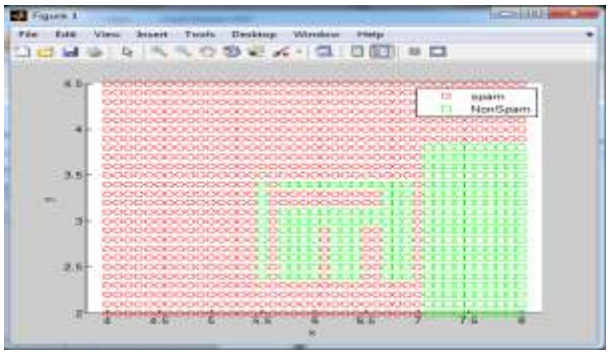

Figure 3.10 Scattering of decision tree based evaluation.

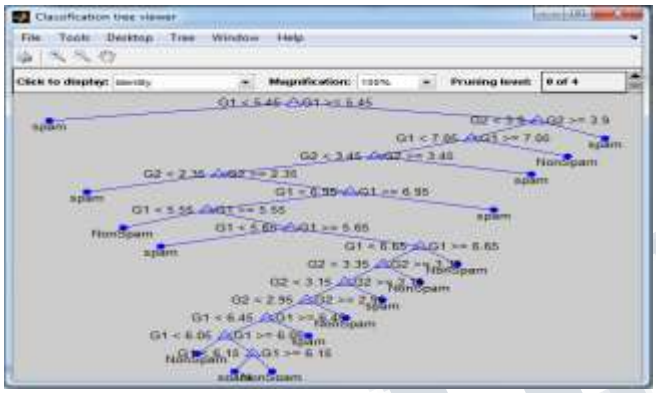

Figure 3.11 General classification of the email dataset in decision tree.

'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'spam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'spam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'spam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'spam'	'NonSpam'
'NonSpam'	'NonSpam'
'spam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'

Figure 3.12 Display of Decision tree classifications against original dataset

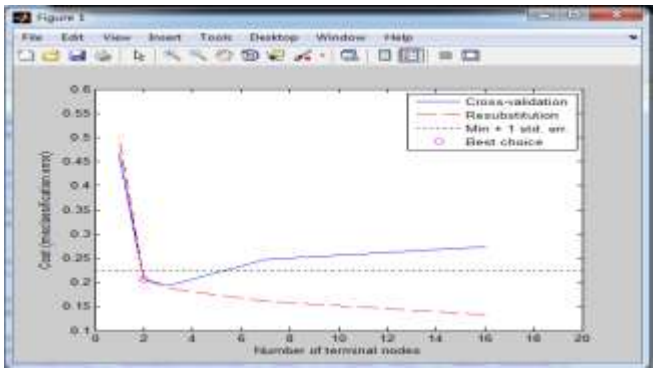

Figure 3.12 plotting the best choice.

In this case the cost of the nodes was calculated and on the basis of this the best choice for the node is determined.

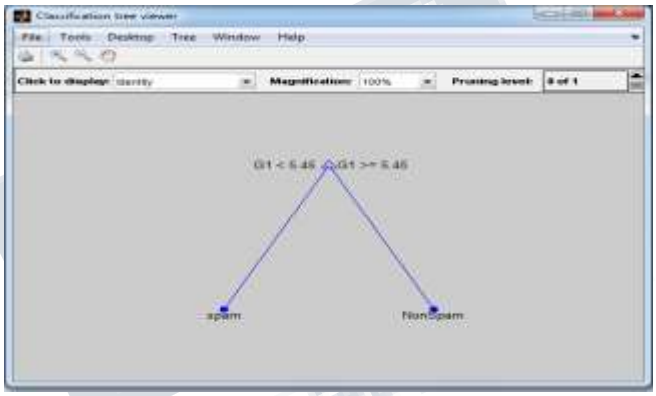

Figure 3.13 Best Level using Decision Tree classification.

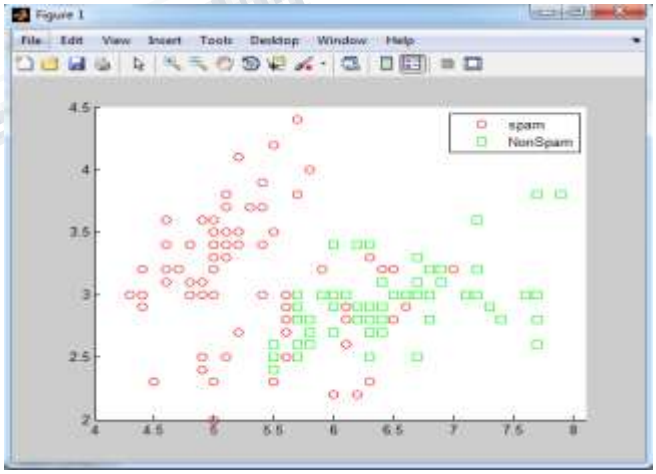

Figure 3.14 Classification plotted using decision tree classifier.

REFERENCES

[1] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk email., AAAI Workshop on Learning for Text Categorization, July 1998, Madison, Wisconsin. AAAI Technical Report WS-98-05.

[2] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos, and P. Stamatopoulos. Learning to Filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. Proceedings ofWorkshop on Machine Learning and Textual Information Access, pages 1-13, year- June-2000.
[3] David Mertz, "Comparing a Half-Dozen Approaches to Eliminating Unwanted Email", August 2002.
[4] Zdziarski, J. A. Ending Spam: Bayesian Content Filtering and The Art of Statistical Language Classification. No Starch Press, San Francisco, CA, USA, 2005.
[5] AhmedKhorsi, "An Overview of Content-Based Spam Filtering Techniques", Informatics 31 (2007) 269-277, 269.
[6] Liu, P. Y., Zhang, L. W., & Zhu, Z. F. (2009). Research on e-mail filtering based on improved Bayesian. Journal of Computers, 4(3), 271-275.
[7] Christina V Karpagavalli S Suganya G,"A Study on Email Spam Filtering Techniques", International Journal of Computer Applicaions12(1):7-9-December-2010 by IJCA Journal , Number1- Article 2.
[8] M. Basavaraju, Dr. R. Prabhakar, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach" ,International Journal of Computer Applications (0975 – 8887) Volume 5– No.4, August 2010.
[9] F.Agayevstreet,Baku,Azerbaijan, Survey on spam Filtering techniques Communication and Network ,Institute of Information technology of Azerbaijan National Academy of sciences, pages 3,153-160 doi:10.4236/cn.2011.33019 published Online August 2011.
[10] V S Kumar , Ravi Kumar," An Efficient Model Of Detection And Filtering Technique Over Malicious And Spam E-Mails" by IJETT Journal Volume-5 Number-1 Year of Publication : 2013.