International Journal of Advanced Research in Engineering Technology & Sciences ISSN: 2394-2819

June-2016 Volume 3, Issue-6

Email:editor@ijarets.org www.ijarets.org

IMPLEMENTATION OF DATA MINING TECHNIQUES FOR DETECTION OF MISCLASSIFICATION ERRORS IN MAILS

Satnam Singh
Ph.D Scholar
Computer Engineering
NIILM University
Kaithal, Haryana

Dr. Pankaj Kumar Verma
Associate Professor
Department of CSE
NIILM University
Kaithal, Haryana

ABSTRACT: As we know that emails are easy to use. Emails are fast and language used in emails is simple can be formal or informal. Message through email delivered at once. There is no paper work while using email. It contains friendly environment and can also have pictures, audio files, video files etc .There is also auto responders in email. Products can be advertised, so that companies can reach a lot of people and can advertise their product in a very short time. But having all these advantages emails have some disadvantages too like emails can carry viruses. Unknown and unwanted people can also send messages called spams. Through emails ones systems can get crashed. Mailbox may get flooded with emails after a certain time so one has to empty it from time to time. 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 by Classification schemes which will save our time and data, by categorizing between spam and non-spam.

1. INTRODUCTION

E-mail is very fast in comparison with the ordinary post and it is easy to use. Emails are fast and language used in emails is simple can be formal or informal. Message through email delivered at once. There is no paper work while using email. It contains friendly environment and can also have pictures, audio files, video files etc .There is also auto responders in email. Products can be advertised, so that companies can reach a lot of people and can advertise their product in a very short time. But having all these advantages emails have some disadvantages too like emails can carry viruses. Unknown and unwanted people can also send messages called spams. Through emails ones systems can get crashed. Mailbox may get flooded with emails after a certain time so one has to empty it from time to time.

1.1 Email Filtering

Email filtering [1-2] 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.

1.2 Classification

Classification is a data mining function that assigns items in a collection to target categories or class. The goal of classification is to accurately predict the target class for each case in the data. For example, a medical researcher wants to analyze breast cancer data to predict which one of three specification treatments a patient should receive. In each of these examples, the data analysis task is classification, where a model or classifier is constructed to predict class (categorical) labels, such as 'safe' or 'risky' for the loan application data; 'yes', 'no' for the marketing data ;or 'treatment A', 'treatment B', 'treatment C' for the medical data. These categories can be represented by discrete values, where the ordering among values has no meaning. Data classification is a two step process [3], 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).

1.3 Classification Algorithms

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[4].

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.

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 original training data into higher dimension. Within this new dimension, it searches for the linear optimal separating hyper plane (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 hyper plane of the classes yield by individual trees (Random Forests is a trademark of Leo Breiman [5] and Adele Cutler for a troupe of choice trees).

2. LITERATURE REVIEW

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 **[6]**. 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 **[7].** In this approach, proper filtering is for the junk (spam) mails and irrelevant mails. So proper filter for these applied on email servers If dynamic nature is used with available spam filters then the results of spam filtering can be unexpectedly better than earlier. So in this research some techniques for improvement of Bayesian filter are discussed. **[8]**. the proposed technique includes the distance between all of the attributes of an email and implemented using open source technology in C language; ling spam corpus dataset was selected for the experiment.

Different performance measures such as the precision, recall, specificity & the accuracy, etc. were observed. K-means clustering algorithm works well for smaller data sets. The work presented in this paper can be further extended & can be tested with different algorithms and varying size of large data sets **[9]**. Proposes the best classifier and better classification approach using different data mining tools using bench mark data set The data set consists of 9324 records and 500 attributes used for training and testing to build the model. In this paper a procedure that can help eliminate unsolicited commercial e-mail,viruses,torjans and worms as well as frauds perpetrated electronically and other undesired and troublesome e-mail. This paper showed analyzing of different supervised classifiers technique using different data mining tools such as weka, rapid miner and support vector machine. This paper showed weka data mining tool give highest accuracy over different data mining tools **[10]**.

3. PROPOSED WORK

My Proposed 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 by Classification schemes which will save our time and data, by categorizing between spam and non-spam. 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. Here I have implemented

4. RESEARCH METHODOLOGY

As the base of our research is to use parallel algorithms, so we will be implementing the linear and quadratic Discriminant analysis, Naïve Bay's algorithm and decision tree, so we will be implementing the standard decision tree algorithm.

- Classification using linear distribution
- Classification plotted using Quadratic Distribution
- Classification using Naive Bayes Gaussian distribution

5. IMPLEMENTATION AND RESULTS

Data mining refers to extracting or "mining" knowledge from large amount of data. It can also be named by "knowledge mining from the data". There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from the databases, knowledge extraction, data pattern analysis, data archaeology and data dredging. Many people treat data mining as a synonym for another popularly used term, knowledge discovery in databases or KDD. Alternatively, data mining is also treated simply as an essential step in the process of knowledge discovery in databases. The fast growing tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. In such situation we become data rich but information poor. Consequentially, important decision are made based on based not on the information-rich data stored in databases but rather than on the decision maker"s intuition, simply because the decision maker does not have the tools to extract the value able knowledge embedded in the vast amounts of data. The dataset for the implementation is taken from the machine learning dataset website "UCI" Repository". The software used for the development of the classification system Weka 3.6 (for the visualization of the dataset) and MATLAB 8 (R2012a). The numeric data is imported to the dataset variable and the class labels are stored in mail group variable.

Table: 1 Dataset information of spam mails from UCI repository

Dataset Characteristics	Multivariate	**Number of**	**Instances**	4601
Attribute Characteristics	Integer, Real	**Number of**	**Attributes**	57

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 occurring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. Here are the definitions of 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.

5.1 MATLAB for linear distribution

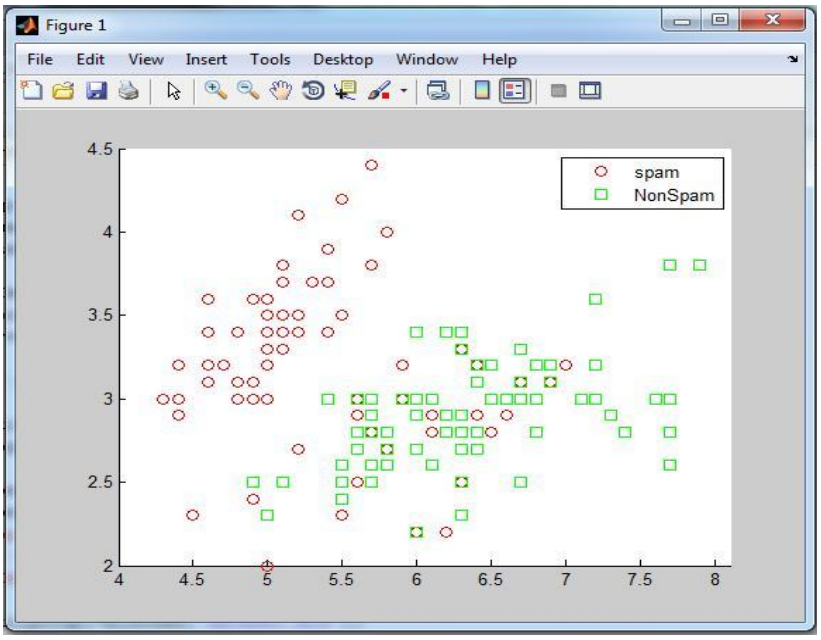

Figure: 1 scattering of the dataset on the basis of the class labels spam and No spam.

This code retrieves the dataset into the workspace and stores the numeric part into dataset variable and class label into mail group. Then the data is scattered and displayed in the graph.

Here N is the number of dataset used= 150 datasets.

The classification is now done on the basis of linear distribution technique. The linear distribution result is displayed against the original class labels and errors or misclassification rate is measured.

The number of mismatch in class labels =28

Therefore the Re substitution error rate: 28/150= 0.1867

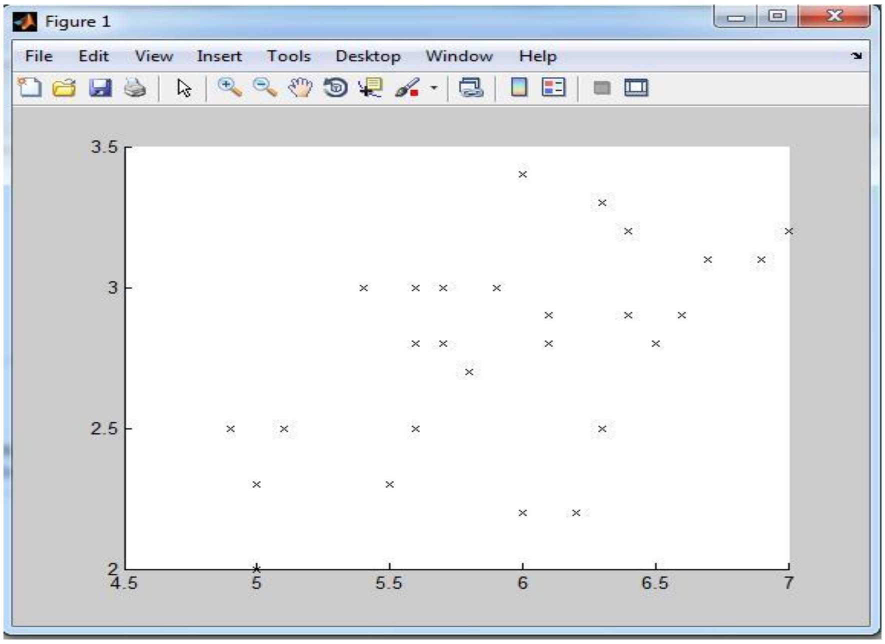

Figure: 2 Misclassification plotted of Spam and Non Spam

The above figure depicts the bad sectors in which the spam and the non-spam mails are merged and we are not able to classify between spam and non-spam. To classify between spam and non spam different classification algorithms are used to detect the misclassification.

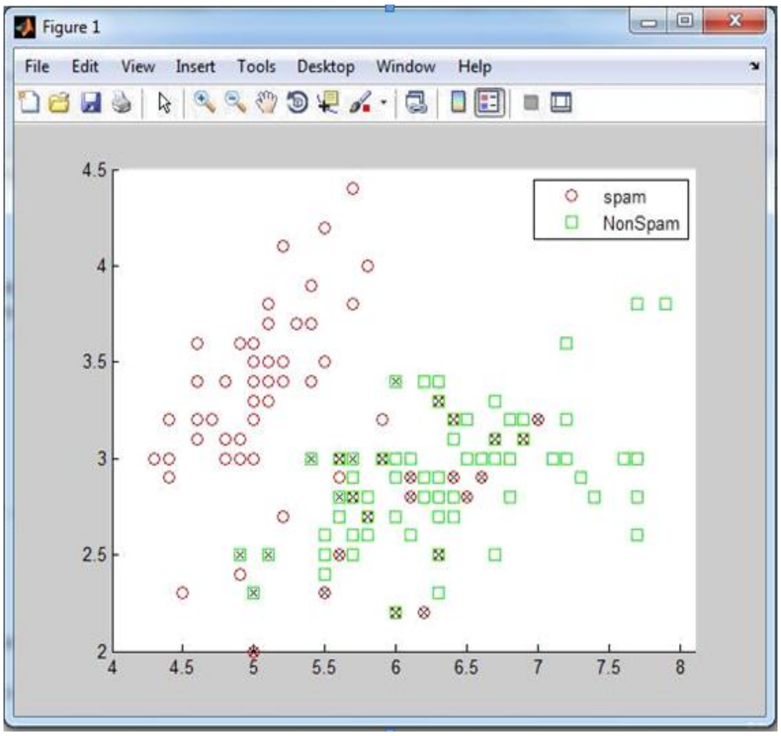

Figure: 3 Misclassification plotted on original scattered class labels

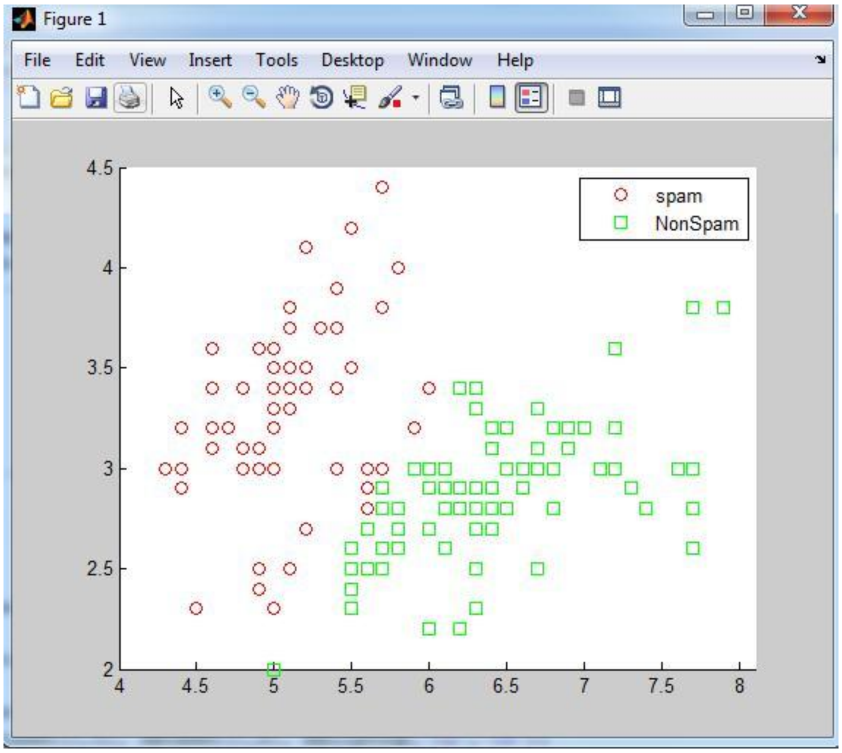

Figure: 4 Classification using Linear distribution algorithm

5.2 MATLAB for quadratic distribution

The total number of misclassification = 25
Therefore the re substitution error rate:
25/150 = 0.1667

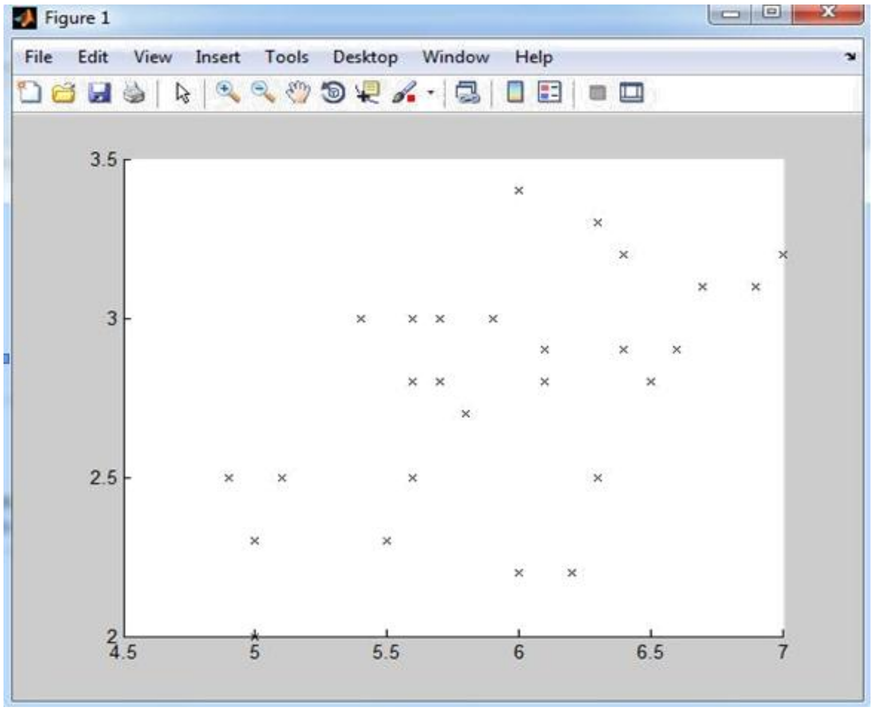

Figure: 5 Misclassification plotted of Spam and Non Spam

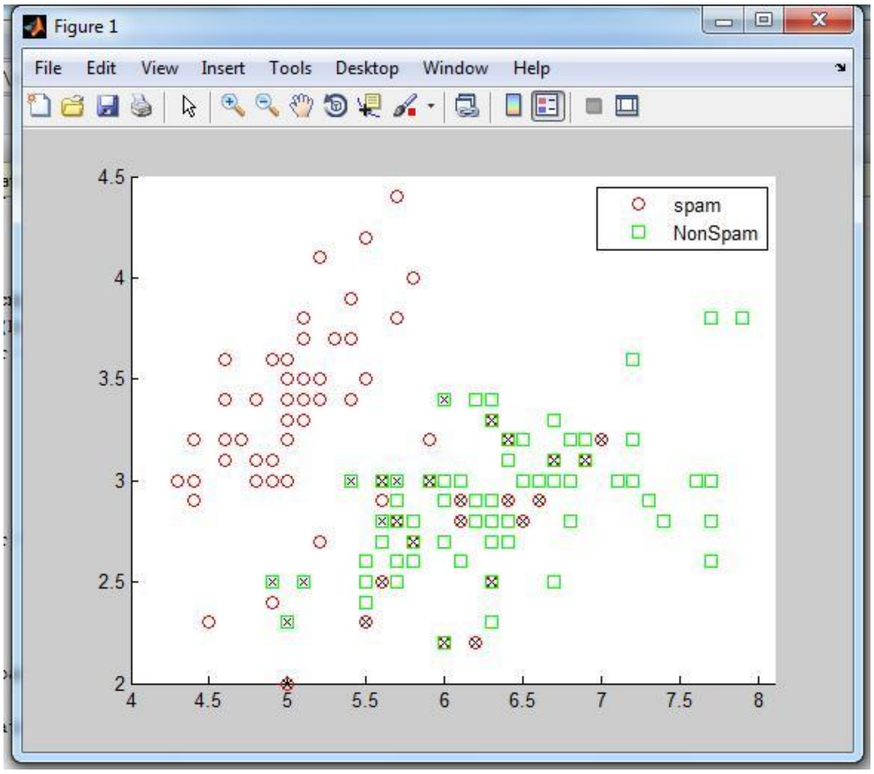

Figure: 6 Misclassification plotted on original scattered class labels

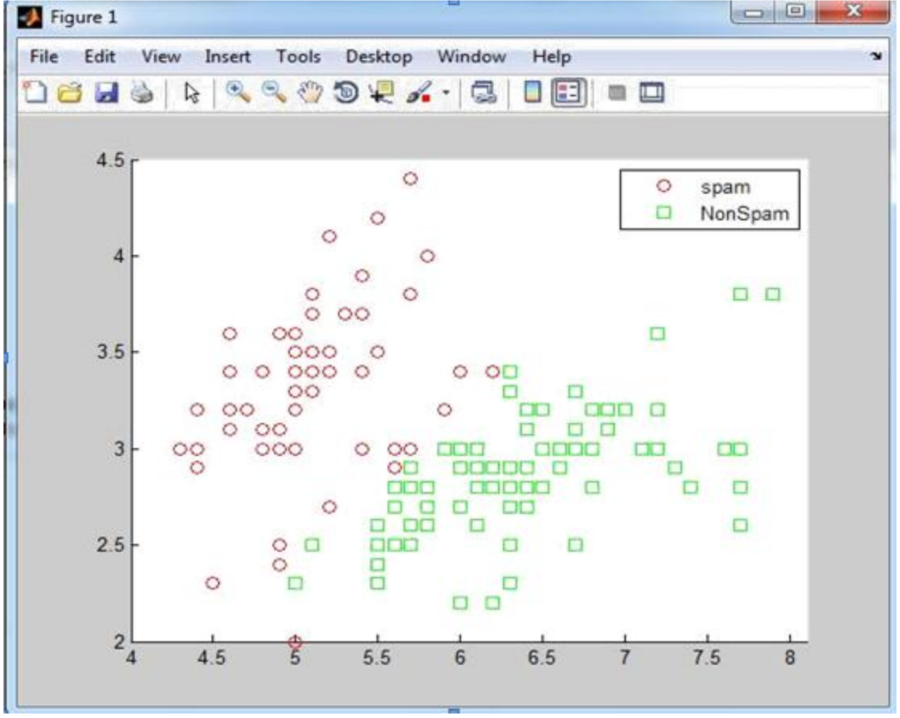

Figure: 7 Classification plotted using Quadratic Distribution

5.3 MATLAB for Naive Bayes Classification

The dataset is now divided into training and testing data for Naive Bayes classification algorithm. The number of test set is 20 and total number of dataset is 150. The Naive Bayes method includes fit and predicts function to evaluate the training set and predicting the class labels on that basis. The cross validation partition allows the dataset to be divided into those parts.

The total number of misclassification in Naive Bayes = 30
The Resubstitution error rate: - 30/150 = 0.200
The cross validation error rate = 0.200

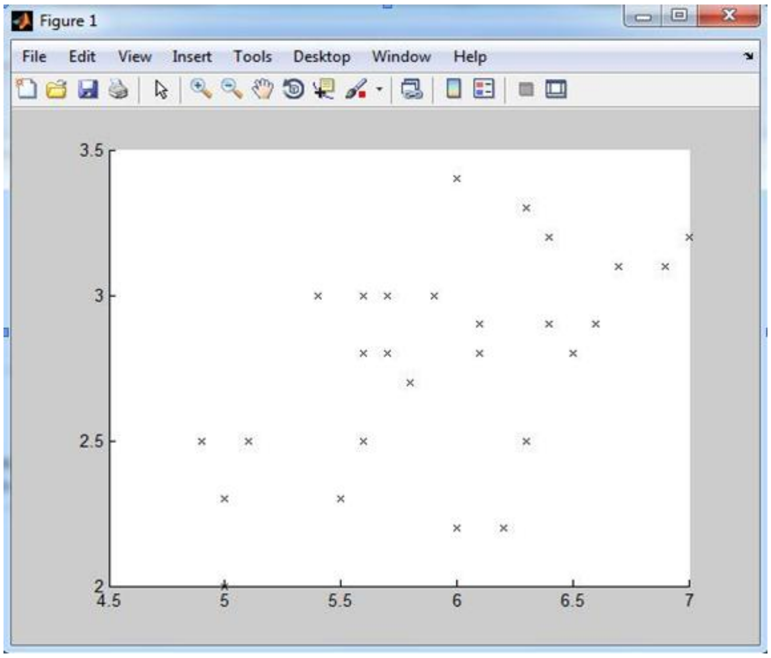

Figure: 8 Misclassification plotted of Spam and Non Spam

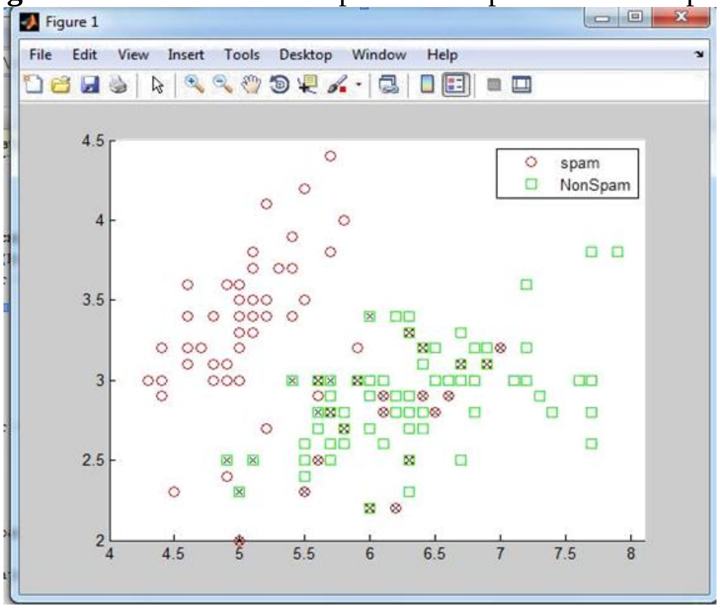

Figure: 9 Misclassification plotted on original scattered class labels

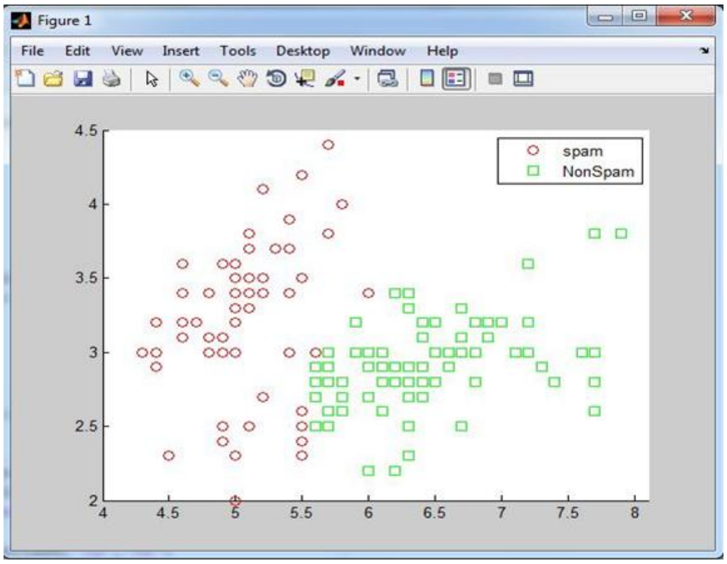

Figure: 10 Classification using Naive Bayes Gaussian distribution

Table: 2 Calculation of cost and seacoast of nodes

Parameters	
Cost	**Secost**
0.2733	0.0362
0.2467	0.0347
0.1933	0.0305
0.2067	0.0305
0.4667	0.0407

Table: 3 Calculation of N term nodes and Resubcost of nodes

Parameters	
N term nodes	**Re sub cost**
16	0.1333
7	0.1600
3	0.1867
2	0.2067
1	0.5000

Therefore the final cost of the best level = cost (bestlevel+1) = 0.2067

Table: 4 Overall Results

CLASSIFICATION	**MISCLASSIFICATION**	**ERROR**
Linear Resubstitution error	28	0.1867
Quadratic Resubstitution error	25	0.1667
NAÏVE BAYES	**MISCLASSIFICATION**	**ERROR**
Gaussian Resubstitution error	30	0.2000
Gaussian Cross validation	30	0.2000

Above calculations and comparison proves that decision tree provides the best results for the classification. Two broad methods are available to estimate the error (misclassification rate) of a classifier. Resubstitution fits a single classifier to the data, and applies this classifier in turn to each data observation. Cross-validation (in leave-one-out form) removes each observation in turn, constructs the classifier, and then computes whether this leave-one-out classifier correctly classifies the deleted observation. Resubstitution typically underestimates classifier error, severely so in many cases. Cross-validation has the advantage of producing an effectively unbiased error estimate, but the estimate is highly variable. In many applications it is not the misclassification rate per se that is of interest, but rather the construction of sets that have the potential to classify or predict. Hence, one needs to rank feature sets based on their performance.

7. CONCLUSION

The work presented by this research is the classification techniques. Therefore, it's a good enterprise solution for filtering. This will optimize the system performance and make some improvements on the previous algorithm. This will give the better results from the previous one. In this paper the filtered mails are further filtered to measure the misclassification using different data mining techniques. This paper shows that the classification using linear and quadratic method is the best classifier. It is easy to interpret and explain the executives. In comparison to random forests are time efficient. Decision tree requires relatively less effort from users for data preparation. For proper visualization and calculation, weka tool and MATLAB has been used. It is often possible to find a simpler tree that performs better than a more complex tree on new data.

REFERENCE

1. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. "*A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization:"* the 1998 workshop (Vol. 62, pp. 98-105).
2. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. *"Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach.arXiv preprint cs/0009009"*(2000).
3. Jason D.M.Rennie "*Practical Concerns Surrounding The Application Of Text Classification To The Problem Of Mail Filtering*", International Journal of Computer Applications (2000).
4. Ian H. Witten, Eibe Frank, *"Data Mining – Practical Mahine Learning Tools and Techniques"* 2nd Edition, Elsevier, (2005).
5. Han, J., Kamber, M., & Pei, J.*"Data mining: concepts and techniques." Morgan Kaufmann* (2000).
6. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July)." the 1998 workshop (Vol. 62, pp. 98-105)". *A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization:"*
7. *Konstantious V. Chandrinos,Constantine D.spyropoulos(2000).To detect the spam Naïve Bayesian is trained automatically"*
8. Rajput, Arjun., & Toshniwal(2008) Adaptive "*Spam Filtering based on Bayesian Algorithm."*
9. López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F., "*A Multiobjective Evolutionary Algorithm for Spam EMail Filtering", Proceedings of 2008 3*rd.
10. Rachna mishra,Ramjeevan Singh Thakur et al (2014)".*An efficient approach for supervised learning algorithms using different data mining tools for spam categorization.Journal IEEE.*