International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-5)

Research Article | **May 2015**

An Analysis of Misclassification Error Detection in Mails Using Data Mining Techniques

Ommera Jan
(C.S.E) Dept, Panchkula Engineering College
Mouli, Haryana, India

Heena Khana
Asst. Professor, Panchkula Engineering College
Mouli, Haryana, India

Abstract—

This research is to classify the filtered data. The main purpose of this research is to reduce the error rate of the data and to improve the accuracy. In the previous techniques of classification there may be some misclassification. But in this research the problem of misclassification is reduced. The work is 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.

Keywords-- Email, filtering, Bayesian filters, Spam, decision tree

I. INTRODUCTION

Email is electronic device. 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. Email messages are usually encoded in the ASCII text. 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 broad casting. 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 have to empty it from time to time.

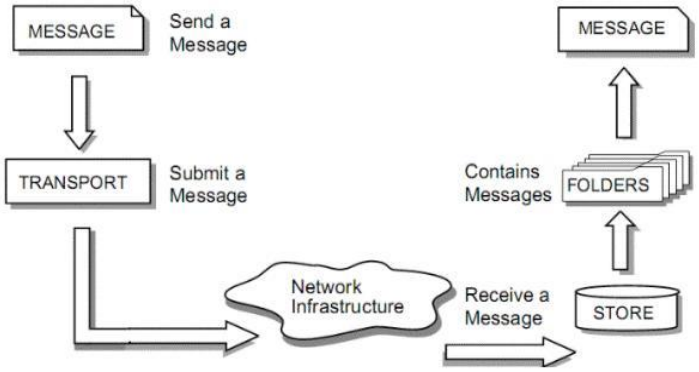

Figure1. Working of email server.

Their messages are modules which are added in the working of the email system. In the processing of the email, SMTP is also used. In shorts, the steps are:

- Message is sent by email client.
- Email server contacted to the recipients email server.
- Username's validity is checked by the email server.
- If valid username is typed, email is sent to the email server of the address.
- When the recipient sings in to his mailing account, he finds his email.

II. PROBLEM FORMULATION

Scope of Study- 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 its classification 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 be 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.

In our base paper [1], it has been declared that random forest algorithm is the best to classify spam and non-spam mails. But there are some advantages of decision tree over random forest. These are as follows:

1. Decision tree is easy to explain and interpret.
2. Decision tree takes less time to be executed than random forest. So it is time efficient too.

These are the reasons why we are implementing decision tree rather than random forest. Except that, the dataset is already filtered, and as random forest is used when there are some complex dataset, there is no need to implement the random forest again.

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.

Objective-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 and the algorithm with least error will be chosen as the best way to filter emails. The steps are:

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

III. RESULTS

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

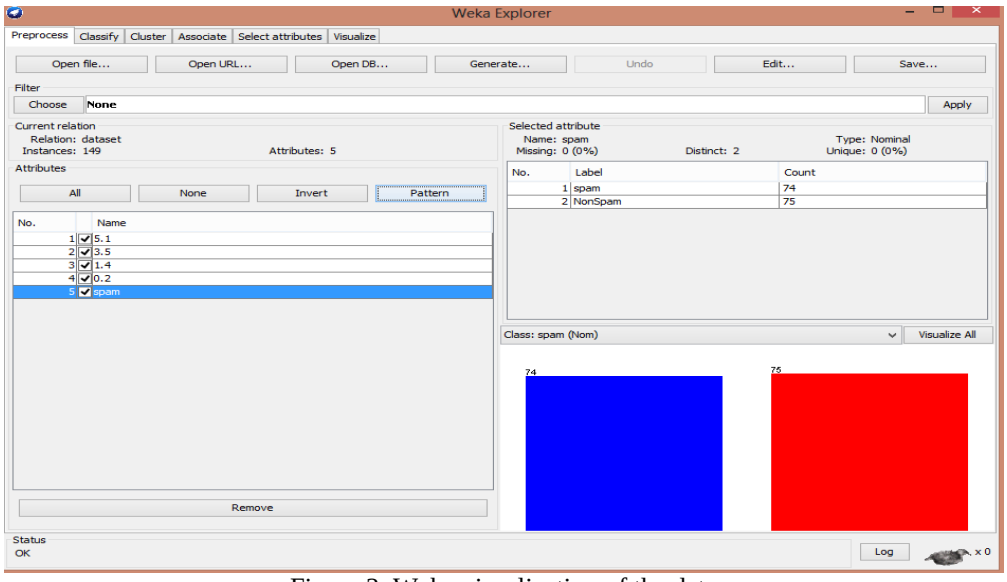

Figure 2. Weka visualization of the data

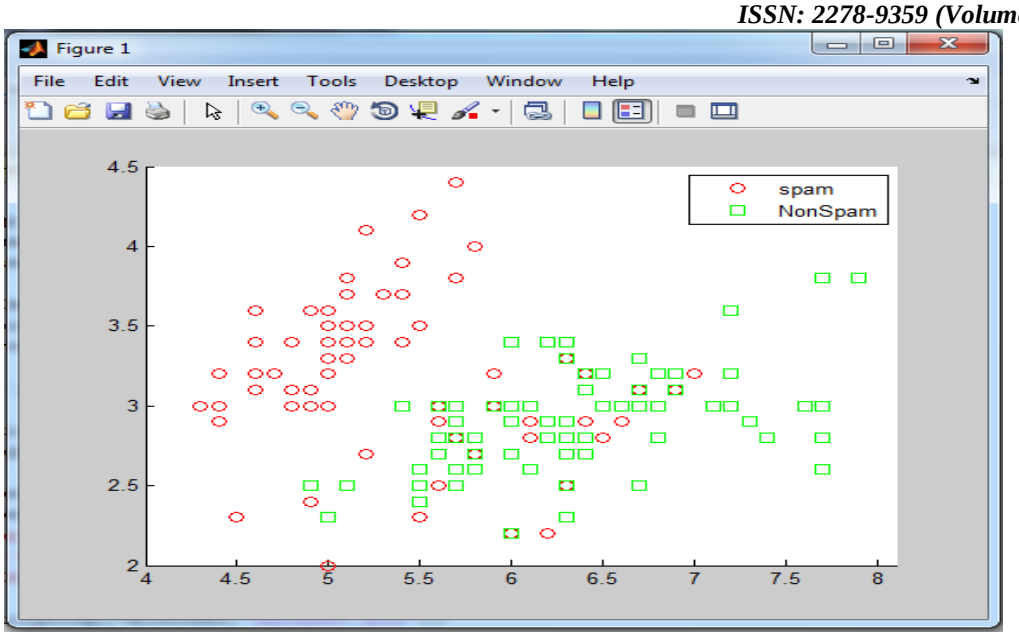

Figure 3.Scattering of the dataset on the basis of the class labels spam and Nonspam

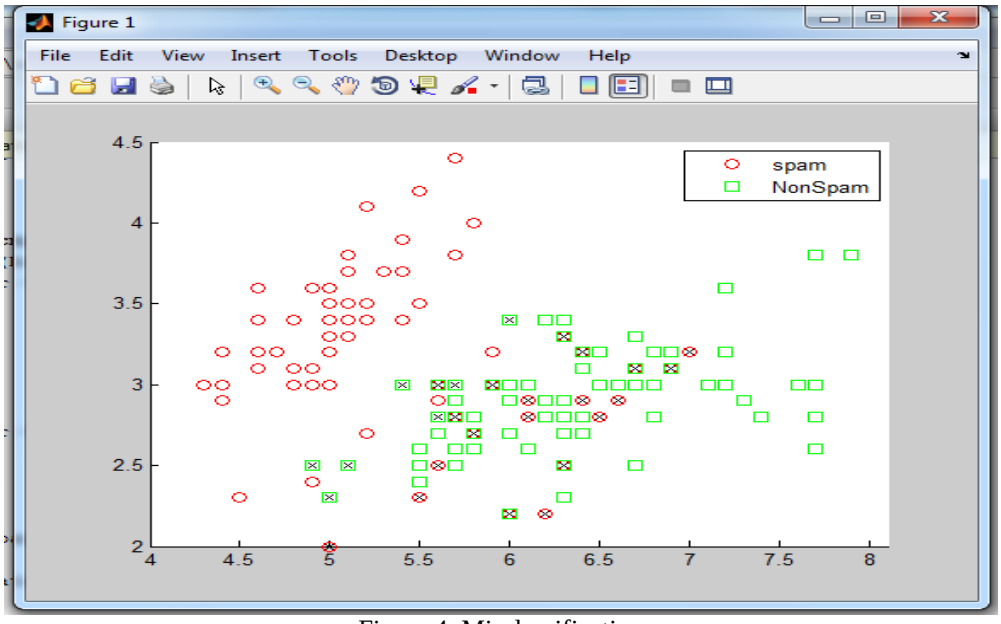

Figure 4. Misclassification

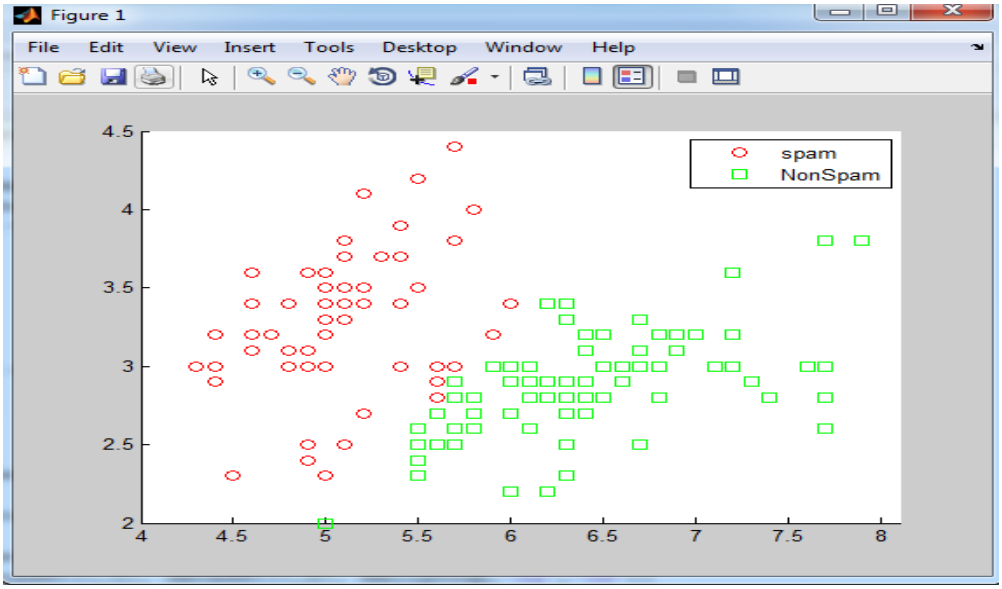

Figure 5. Classification using linear distribution

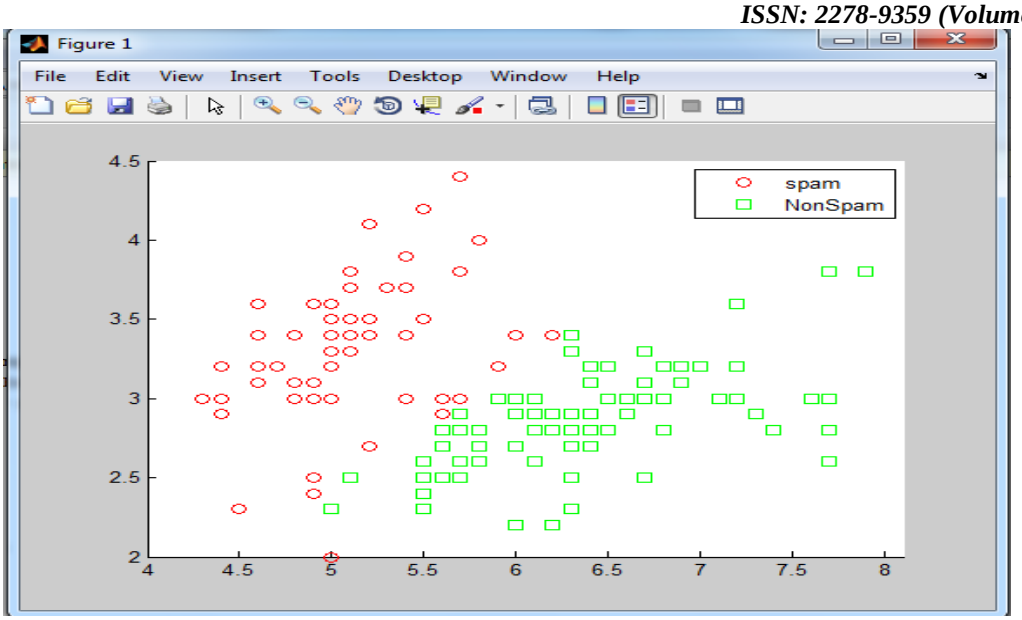

Figure 6.Classification plotted using Quadratic Distribution

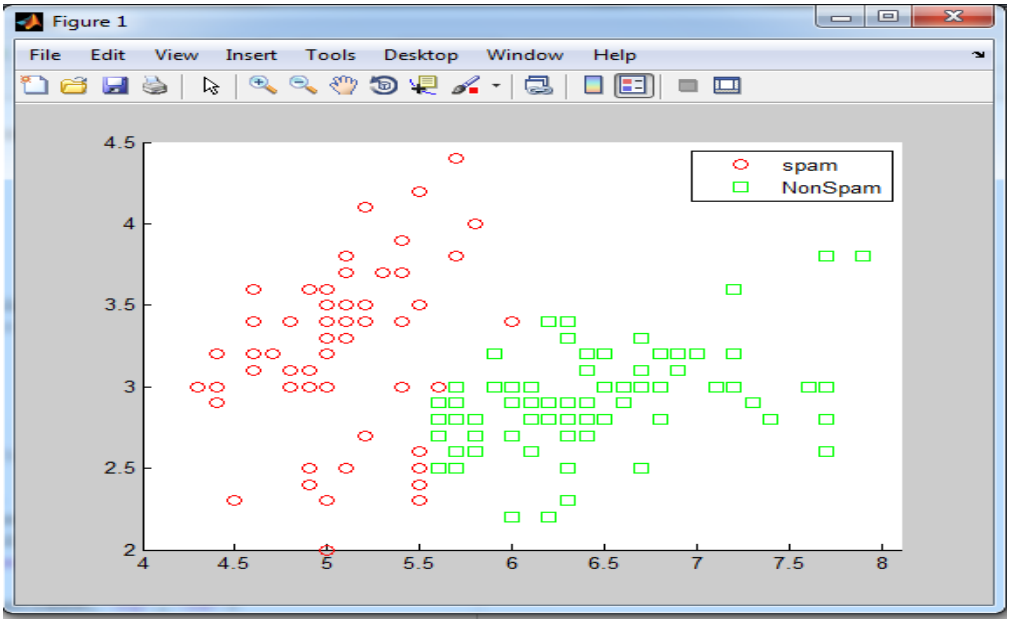

Figure 7. Classification using Naive Bayes Gaussian distribution

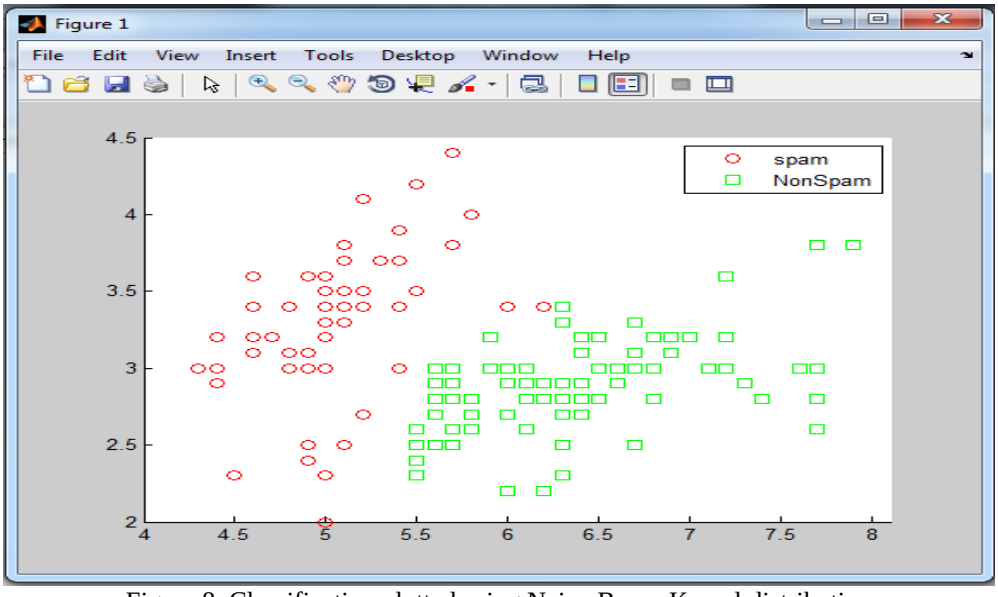

Figure 8. Classification plotted using Naive Bayes Kernel distribution

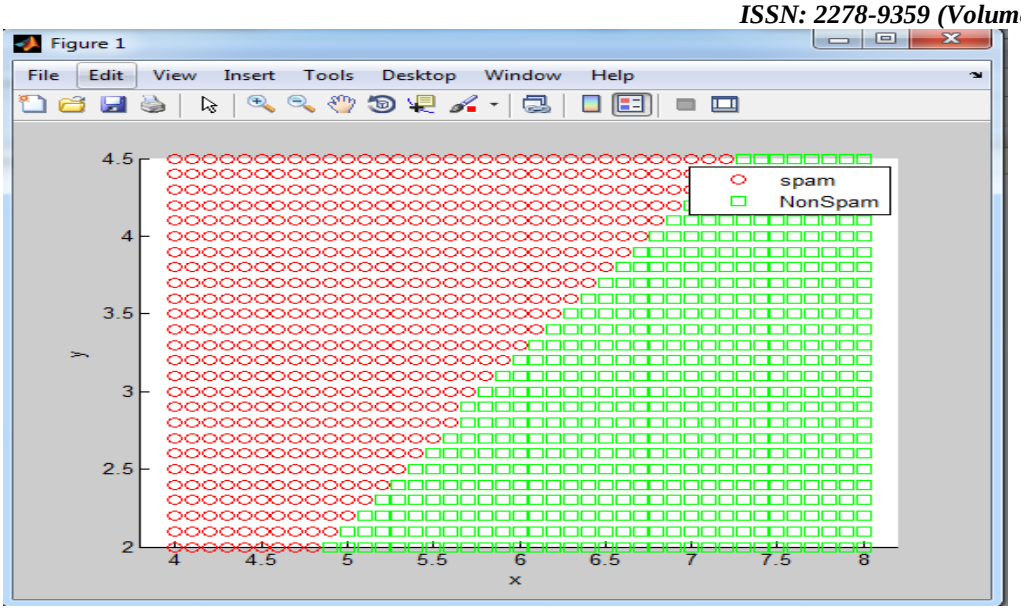

Figure 9.Scattering of mesh grid for x and y axis

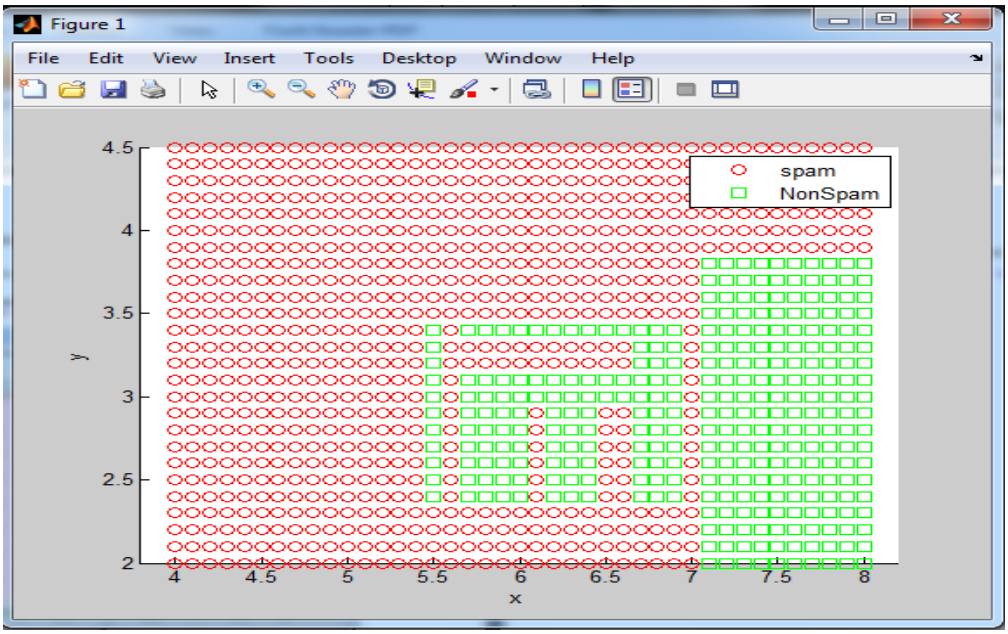

Figure 10. Scattering of decision tree based evaluation

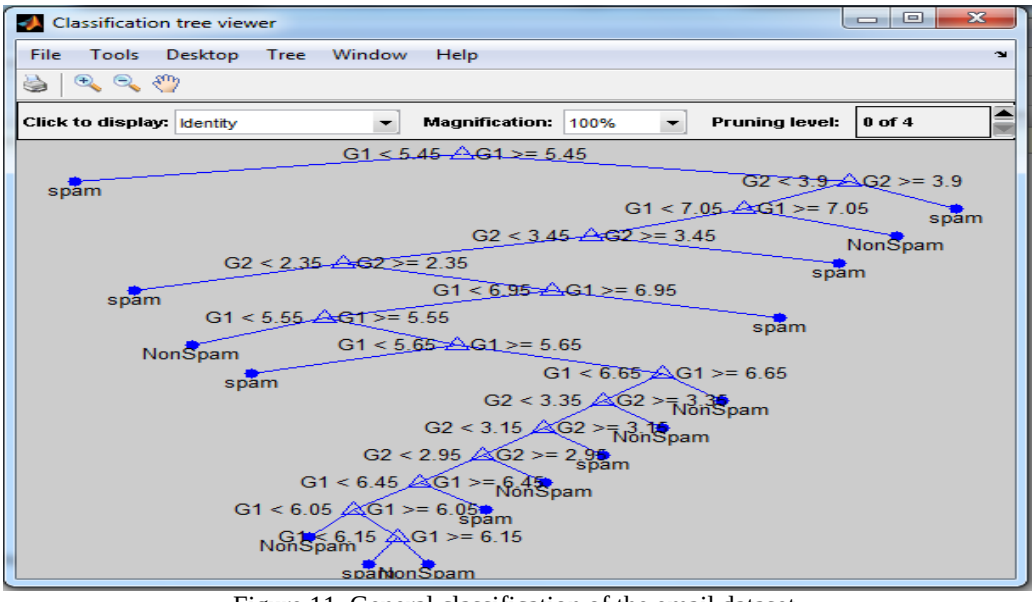

Figure 11. General classification of the email dataset.

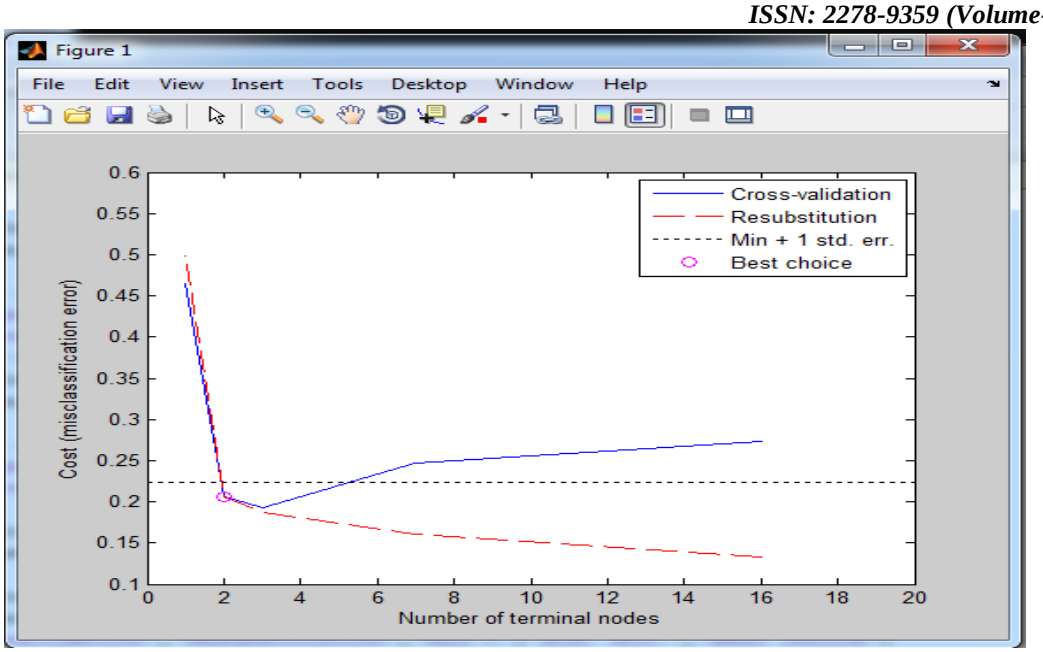

Figure 12. plotting the best choice

Table2.Calculation of cost and secost of nodes

cost	secost
0.2733	0.0362
0.2467	0.0347
0.1933	0.0305
0.2067	0.0305
0.4667	0.0407

Table3.Calculation of ntermnodes and resubcost of nodes

Ntermnodes	resubcost
16	0.1333
7	0.1600
3	0.1867
2	0.2067
1	0.5000

Best level=3

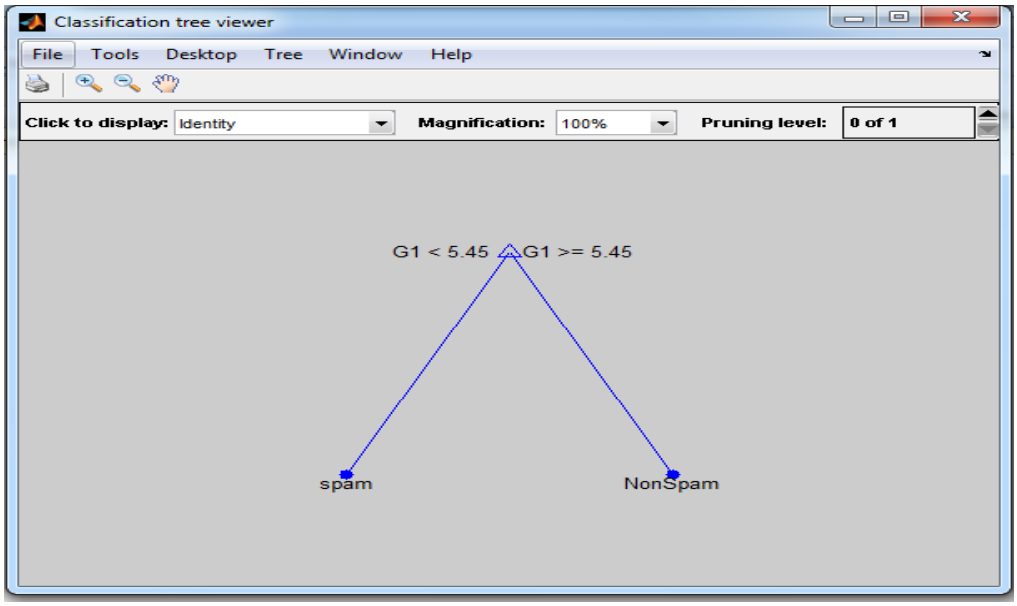

Figure 13. Best Level using Decision Tree classification

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

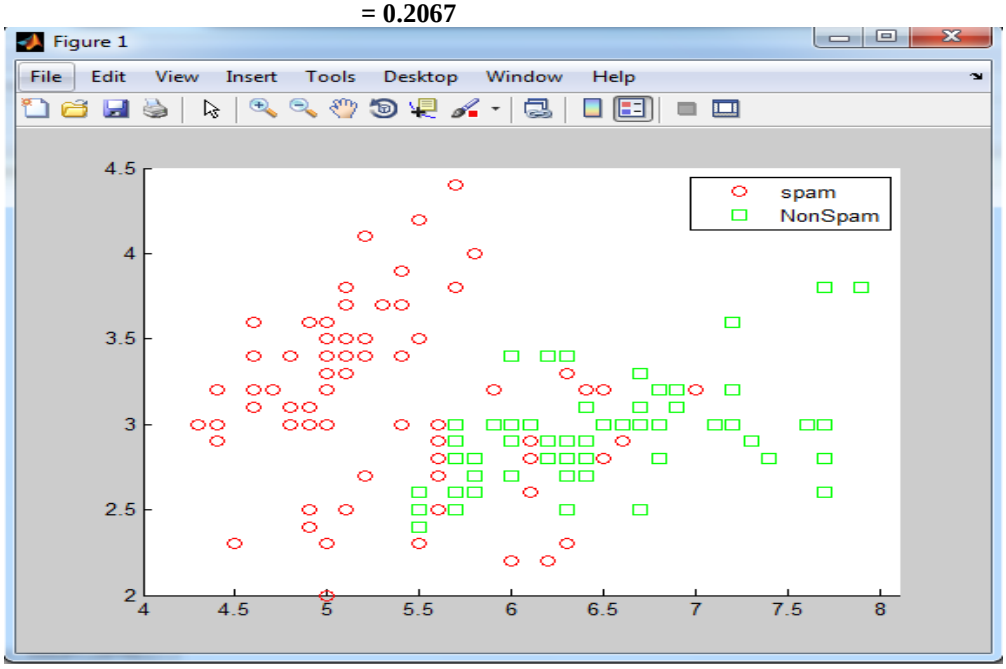

Figure 14. Classification plotted using decision tree classifier

Table 4.Comparision of data mining techniques

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 Resubstitution error	30	0.2000
Kernel distribution resubstitution error	28	0.1867
Kernel distribution cross validation resubstitution error	28	0.1933
DECISION TREE	**MISCLASSIFICATION**	**ERROR**
Resubstitution error	20	0.1333
Cross validation error	20	0.2533

Above calculations and comparison proves that decision tree provides the best results for the classification.

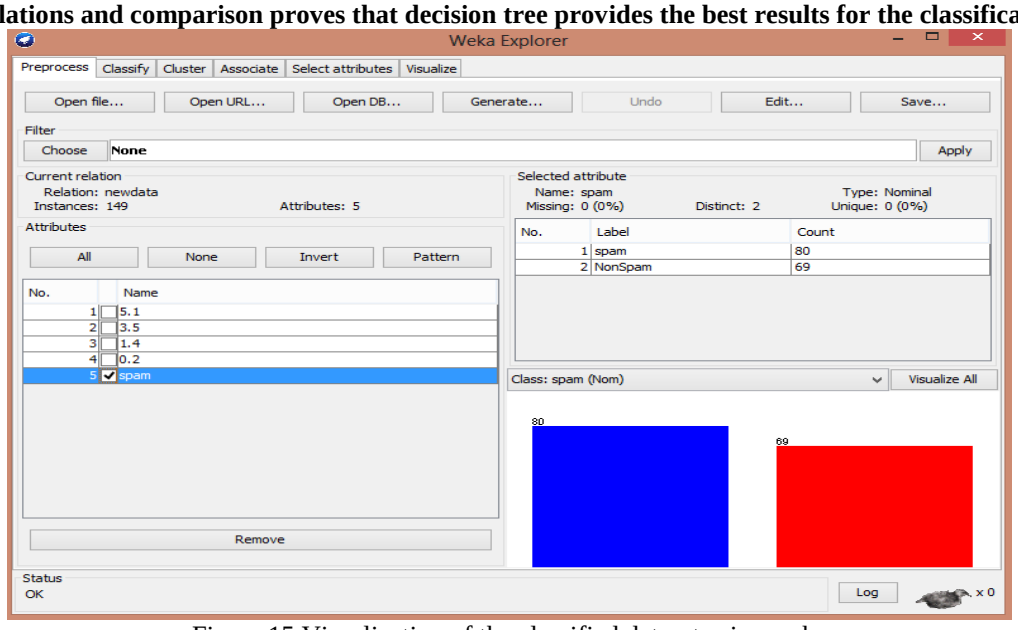

Figure 15.Visualization of the classified dataset using weka

V. CONCLUSIONS AND FUTURE SCOPE

In this paper the filtered mails are further filtered to measure the misclassification using different data mining techniques. This paper shows that the decision tree 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.As a future work of our research, we can tune the parameters of our research with neural network. We can also expand the limit of our dataset while using neural network to extract more accuracy from our results.

ACKNOWLEDGMENT

Working on this thesis **of Analysis of Misclassification Error Detection in Mail Using Data Mining Techniques** provided a unique experience and analysis, I feel great pleasure and privilege in working over this research. I am deeply indebted to **"Panchkula Engg. College"** for the invaluable guidance, support and motivation for the many other aids without which it would have been impossible to complete this project.

I have no words to express my deep sense of gratitude for Heena Khanna (Mentor) for her enlightening guidance, directive encouragement, suggestions and constructive criticism for always listening to our problems and helping us out with their full cooperation.

Last but not the least Father Haji Bashir Ahmad Dar , Mother Jahan Ara, Brothers Nisar Ahmad Mir, Ommer Bashir and friends who have given me that much strength to keep moving on forward every time.we are greatly thankful to them and have no words to express my gratitude to them.

REFERENCES

[1] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.

[2] Basavaraju, M., & Prabhakar, R. (2010). A novel method of spam mail detection using text based clustering approach. International Journal of Computer Applications, 5(4).

[3] Hovold, Johan. (2005, July). Naive bayes spam filtering using word-position-based attributes. In Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS 2005).

[4] Jin, X., Wang, L., Lu, Y., & Shi, C. (2003). MC-tree: Dynamic index structure for partially clustered multi-dimensional database. Tsinghua Science and Technology, 8(2), 174-180.

[5] 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.

[6] Rajput, Arjun., & Toshniwal, D. Adaptive Spam Filtering based on Bayesian Algorithm.

[7] Rennie, J. (2000, August). ifile: An application of machine learning to e-mail filtering. In Proc. KDD 2000 Workshop on Text Mining, Boston, MA.

[8] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).

[9] Song, Y., Kołcz, A., & Giles, C. L. (2009). Better Naive Bayes classification for high-precision spam detection. Software: Practice and Experience, 39(11), 1003-1024.

[10] Prajwala T R(2015).A comparative study on decision tree and random forest using R tool.International journal of advanced research in computer and communication engineering.(vol.4 196-1

[11] Christina V et al(2010).Problems associated with spam and spam filtering methods.

[12] Konstantious V. Chandrinos,Constantine D.spyropoulos(2000).To detect the spam Naïve Bayesian is trained automatically.

[13] Xiaoming JIN,Yuchang LU et al (2003).Indexing problem in dataset composed of partially clustered data.

[14] 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.

[15] Jehad Ali,Rehanullah khan,Nasir Ahmad,Imran Maqsood(Sep 2012).Random Forests and Decision trees .Journal IJCSI.

