ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Spam Detection Using Data Mining Tool In Matlab

Amandeep Kaur[1], Jatinder Kaur[2]

Student, ECE, CGCTC Jhanjeri, Mohali, India [1]

Assistant Professor ECE, CGCTC Jhanjeri, Mohali, India[2]

Abstract: Our research is focused on distinguish between spam and non spam. The whole procedure is focused on reducing error rate of data being misclassified. Rather than the previous researches where there were issues of classification error, we are going to modify the classification techniques through which better results and minimal error rates are found. This will augment system performance too.

Keywords: Spam detection, Filtering, Decision tree algorithm, Naive Bayes algorithm.

I. INTRODUCTION

E-mail is the method which exchanges the digital message from its source to destination. Email has many advantages but it has disadvantages too like spam means unwanted and unknown people can send message. So that's why email filtering is needed. Filtering means systematize e-mail according to its exact criteria. There are many types of filtering like blacklist filtering, white list filtering, word based filtering, heuristic filtering and Bayesian filtering but Bayesian filtering is the powerful technique and also the bright solution to fight with spam mails nowadays. Email filtering provides many benefits like:
Deal with the service, Improve efficiency, Reduce communication load, Avoid investment, Improve reliability, Increase safety measures and Mitigates liability.
Data Mining Data mining means extracting or "mining" knowledge from large amount of data There are many other terms carrying a similar or slightly different meaning to data mining, such as data pattern analysis, data archaeology and data dredging and data mining is also used as the term knowledge discovery.

II. PROBLEM FORMULATION

Many researchers have done work on spam detection. Previously, spam classification is done on many datasets by using different algorithm and it was found that Random Forest algorithm is best suitable for the same. But there are some disadvantages related to this algorithm. These are:

1. Long hierarchal tree may make the algorithm slow for real-time prediction.
2. This algorithm is not suitable for less number of dataset due to longer execution time.
3. Hard to understand

I provide the improved version by reducing the misclassification. In this work a proposed method is used to vanish above problem.
This method includes two algorithms. These algorithms are:
1. Naive Bayes Classifier
2. Decision tree

III. RESEARCH METHODOLOGY

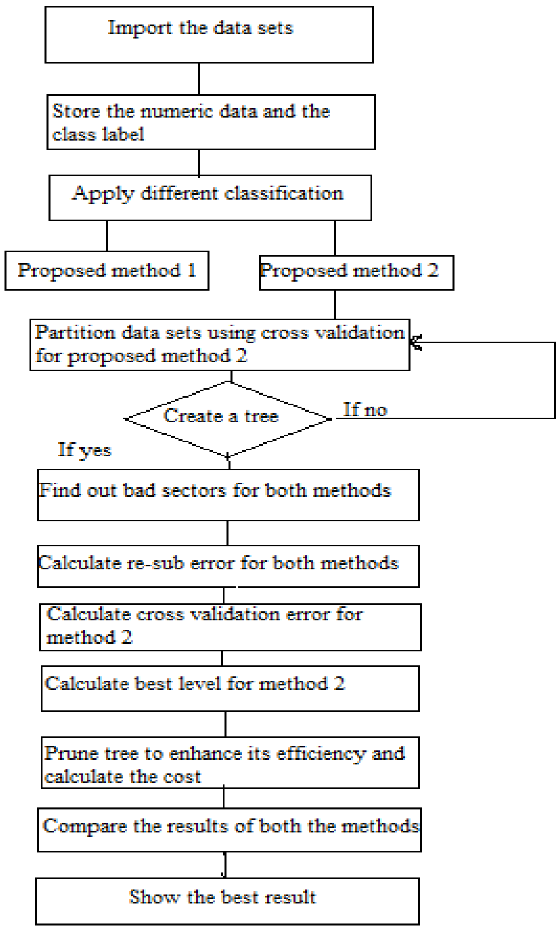

Figure 1 Flow chart of methodology

IV. METHODOLOGY STEPS

Step1 Import the data sets
Step2 Apply different classifications
Step3 Find out the bad sectors for both methods
Step4 Calculate the re substitution error for both methods
Step 5 Calculation cross validation error for method 2 and calculate the best level then calculate the cost

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Step 6 Compare the results of both the methods
Step 7 Show the best result

V. RESULTS AND DISCUSSION

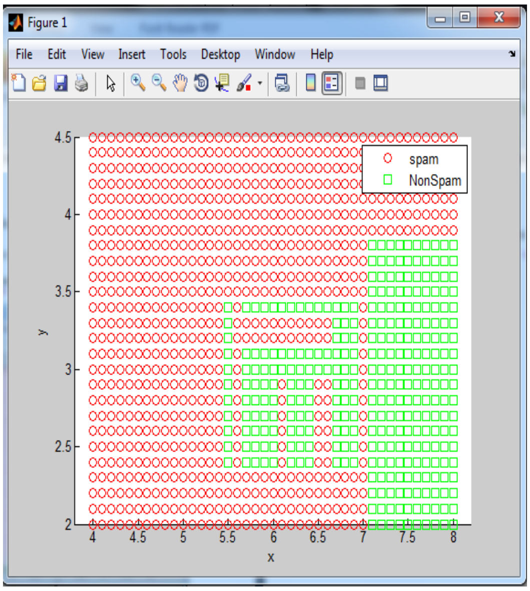

Figure 2.1 Decision tree based evaluation

The evaluation of the tree results into set of variables Grp name = 1066 node = 1066 when dataset = 150 and fig 4.1 depicts the decision tree based evaluation which shows the spam and non spam mails.

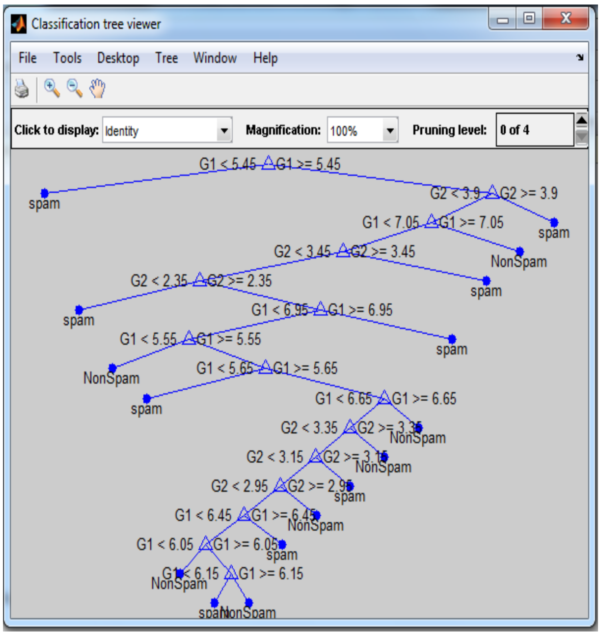

Figure 2.2 General classification of the email dataset

The above figure shows the general classification of datasets and also shows which one is spam or non spam.

Number of misclassification = 20
Re substitution error rate = 20/150
= 0.133
Cross validation error rate = 0.2533

'NonSpam'	'NonSpam'
'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'
'NonSpam'	'NonSpam'
'NonSpam'	'NonSpam'

Table3. Display of Decision tree classification against original dataset

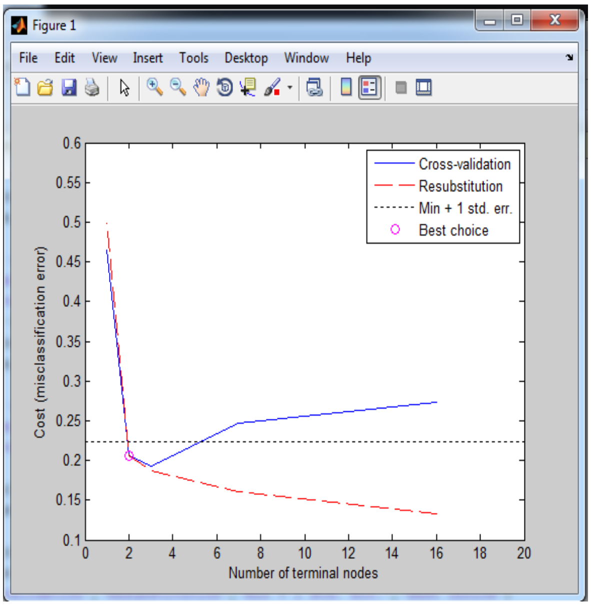

Figure 2.3 plotting the best choice

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

cost =	Se cost =
0.2733	0.0362
0.2467	0.0347
0.1933	0.0305
0.2067	0.0306
0.4667	0.0407
N term nodes =	Re sub cost =
16	0.1333
7	0.1600
3	0.1867
2	0.2067
1	0.5000

Best level =3

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

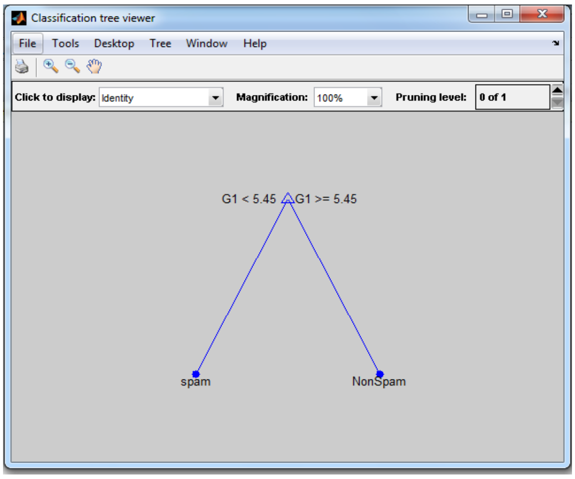

Figure 2.4 Best Level using Decision Tree classification

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

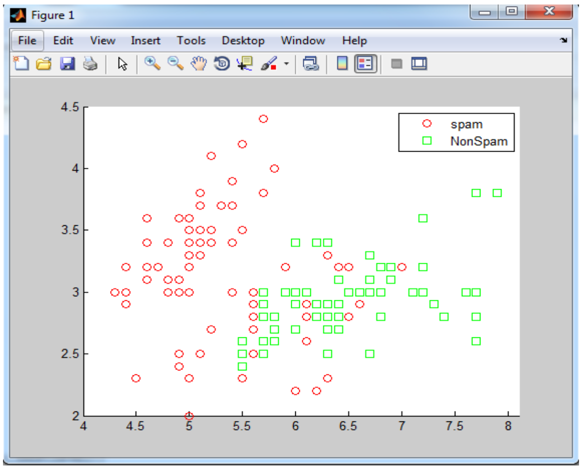

Figure 2.5 Classification plotted using decision tree classifier

The above figure illustrates the improved version in which spam and non spam e-mails are filtered.

VI. CONCLUSION

In our research, we have focused our work on further filtration of email data. We have calculated the linear re-substitution error, quadratic re-substitution error and cross validation error and compared them. We have implemented Naïve Bayes algorithm as well as decision tree algorithm. In previous research, author had worked on Random Forest algorithm. But there are some disadvantages of Random Forest because of which, we have worked on decision tree to diminish the limitations. We have successfully found out the misclassified mails and compared all of them.

REFERENCES

[1] V .Christina, S .Karpagavalli,G.Suganya Email Spam Filtering using Supervised Machine Learning Techniques International Journal on Computer Science and Engineering Vol. 02, No. 09, 2010, 3126-3129.
[2] Hovold , J(2005,july) naïve bayes spam filtering using word-position-based attributes. In proceedings of the 2nd conference on Email and Anti-spam(CEAS 2005).
[3] Liu Pei-yu, Zhang Li-wei, Zhu Zhen-fang Research on E-mail Filtering Based On Improved Bayesian JOURNAL OF COMPUTERS, VOL. 4, NO. 3, MARCH 2009 271
[4] Shangguang Wang, Member, IEEE, Zibin Zheng, Member, IEEE, Zhen ping Wu, Member, IEEE, Fangchun Yang, Member, IEEE, Michael R. Lyu, Fellow, IEEE Reputation Measurement and Malicious Feedback Rating Prevention in Web Service Recommendation Systems IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. , NO. , MARCH 2014
[5] Rachna Mishra, Ramjeevan Singh Thakur,An efficient approach for supervised learning algorithm using different data mining tools for spam categorization, 2014 Fourth International Conference on Communication Systems and Network Technologies
[6] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos and Constantine D. Spyropoulos (2000)Software and Knowledge Engineering Laboratory Institute of Informatics and Telecommunications An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages
[7] S. Wang, Z. Zheng, Q. Sun, H. Zou, and F. Yang. Evaluating Proceedings of the IEEE International Conference on Services Computing (SCC'11), pages 192-199, 2011.[14]
[8] S. Ries and E. Aitenbichler. Limiting Sybil Attacks on Bayesian Trust Models in Open SOA Environments. In Proceedings of the 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (UIC-ATC'09), pages 178-183, 2009
[9] G.Santhi1, S. MariaWenisch2 and Dr. P. Sengutuvan A Content Based Classification of Spam Mails with Fuzzy Word Ranking IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 2, May 2013 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784 www.IJCSI.org
[10] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In Proceedings of the 2th ACM conference on Electronic Commerce (EC'00), pages 150-157, 2000.
[11] S. Nepal, Z. Malik, and A. Bouguettaya. Reputation Propagation in Composite Services. In Proceedings of the IEEE International Conference on Web Services (ICWS'09), pages 295-302, 2009.
[12] Data, C. H. D. (2010). Data Mining: Concepts and Techniques.
[13] Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan Kaufmann.
[14] 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. Ar Xiv preprint cs/0009009.
[15] 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.