iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** IV **Month of publication:** April 2017

DOI: http://doi.org/10.22214/ijraset.2017.4213

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Performance Evaluation of Data Mining based Classifier for Classification of Spam E-Mail

Manish Kumar Sahu[1], A. K. Shrivastava[2]
[1]Dept. of IT, [2]Dept. Of Physics, Dr. C. V. Raman University Kota Bilaspur (C.G.)

Abstract: ***E-mail is one of the important and economical communication media to transfer the information from one person to others. Due to increase number of E-mails resulted drastic increases spam E-mail. In this research work, we have used various classification techniques to classification of spam E-mail and non spam E-mails. The experiment done in Tanagra data mining tool. We have recommended the Multilayer perceptron (MLP) as a best classifier for classification of spam which gives 93.15% accuracy with 10-fold cross validation.***
Keywords: Classification, Multilayer Perceptron (MLP), Cross Validation.

I. INTRODUCTION

Now days the use of Internet is increasing every day to access information in the world wide web. In every organization like bank, insurance, industries have large volume of data. Due to increase number of E-mail users, spam E-mail play very serious problem for E-mail users. Spam e-mail is not necessary to harmful for users, it also wastage the storage space on mail box. To secure information and avoid the unnecessary wastage space in mail box, classification play very important role. Classification is one of the important data mining techniques to classify the spam and non spam E-mail. Many authors have worked in the field of spam E-mail classification. There are various authors have worked in the field of classification of spam e-mail., A. Sabri Taha et al. (2010) have proposed ensemble of continuous learning approach (CLA) and Artificial Neural Network (ANN) (CLA_ANN) for classification of spam e-mail and used spam assassin public corpus data set. Y. Liu et al. (2011) [3] have suggested Word Sequence Kernel based on the Dependent measure (PDWSK) model for spam filtering and compared the result of PDWSK model with other SVM under different kernel functions model. The proposed model gives better results to others as 93.64% of precision, 92.21% of recall and 92.92% of F- measures. L. Varghehese, L. et al. (2012) [4] have also proposed a new hybrid model using VSM and Racchio for classification of spam e-mail. They have compared the result of both the techniques in terms of sensitivity, specificity, precision and F-measures, the proposed model suggested by author outperform the VSM. M. Awad [5] have suggested a new proposed hybrid approach that is combination of RBF Neural Network and Partial Swarm Optimization (RBFNN-PSO) for classification of spam e-mail. R. Sharma et al. (2016) [6] have suggested and compare the performance of RBF and SVM techniques for classification of phishing e-mail. The proposed RBF gives better performance compare to SVM technique.

II. METHOD AND MATERIALS

Tools and techniques are very important role in fields of every research area. This experiment carried out using Tanagra data mining software and various data mining based classification techniques. We have used various data mining based classification techniques like decision ,multilayer perceptron (MLP), Radial basis function (RBF) and support vector machine (SVM) for classification of spam E-mail data set. Decision tree (Jiawei Han and Micheline Kamber, 2006) is most popular and powerful classification techniques in which in the training stage a tree like structure is formed where each non-leaf node is decision node which splits according to the features of training data while leaf node represent class node, Once the decision tree is formed, unknown samples can be presented to the root node of decision tree and ultimately reaches to the class node to classify the sample as one of the target class. In this work , we have used decision tree algorithm (A. K. Puj1193ari,2001) as C4.5 , CART and ID3. We have also used various learning techniques like MLP, RBF and SVM (A. K. Pujari, 2001) for classification of spam E-mail.

In this research work, we have used spam email data set collected from UCI repository. Spam E-mail data set contain 57 features with 4601 samples in all, out of which 1813(37.4%) samples are related to spam while rest of the samples i.e. 2788 (60.6%) are related to non-spam.

This research work used open source Tanagra data mining software for analyzing of data. This tool is used in various application of data mining like classification, prediction, clustering and association rule mining.

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

III. RESULTS AND DISCUSSION

In this research work, we have used Tanagra data mining software in window environment with i5 system. We have used spam e-mail data set that is applied into various data mining techniques like C4.5, CART, ID3, MLP, SVM and RBF for analysis of data and identify and classification of spam and non-spam e-mail. Table 1 to table 6 shows that the confusion matrix and error rate of various models. We have calculated the various performance measures of these models using confusion matrix. From the tables , it is clear that error rate of MLP algorithm gives minimum error rate as 0.0685. All the experiment done in 10-fold cross validation. Table 7 shows that performance measures like accuracy, sensitivity and specificity. MLP gives better accuracy and specificity as 93.15% and 95.37% respectively. C4.5 gives best sensitivity as 89.96%. Fig. 1 shows that performance measures of various models.

Table 1: Analysis of C4.5 model with 10-fold cross validation

Error rate			0.0846			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		spam	non_spam	Sum
spam	0.8996	0.1126	spam	1631	182	1813
non_spam	0.9257	0.0659	non_spam	207	2580	2787
			Sum	1838	2762	4600

Table 2: Analysis of CART model with 10-fold cross validation

Error rate			0.0900			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		spam	non_spam	Sum
spam	0.8538	0.0878	spam	1548	265	1813
non_spam	0.9465	0.0913	non_spam	149	2638	2787
			Sum	1697	2903	4600

Table 3: Analysis of ID3 model with 10-fold cross validation

Error rate			0.1033			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		spam	non_spam	Sum
spam	0.8737	0.1344	spam	1584	229	1813
non_spam	0.9117	0.0827	non_spam	246	2541	2787
			Sum	1830	2770	4600

Table 4: Analysis of MLP model with 10-fold cross validation

Error rate			0.0685			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		spam	non_spam	Sum
spam	0.8974	0.0735	spam	1627	186	1813
non_spam	0.9537	0.0654	non_spam	129	2658	2787
			Sum	1756	2844	4600

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Table 5: Analysis of SVM model with 10-fold cross validation

Error rate	0.1002
Values prediction	Confusion matrix

Value	Recall	1-Precision
spam	0.8224	0.0853
non_spam	0.9501	0.1084

	spam	non_spam	Sum
spam	1491	322	1813
non_spam	139	2648	2787
Sum	1630	2970	4600

Table 6: Analysis of RBF model with 10-fold cross validation

Error rate	0.1720
Values prediction	Confusion matrix

Value	Recall	1-Precision
spam	0.8991	0.2717
non_spam	0.7818	0.0775

	spam	non_spam	Sum
spam	1630	183	1813
non_spam	608	2179	2787
Sum	2238	2362	4600

Table 7: Performance measures of various models

Model	Accuracy	Sensitivity	Specificity
C4.5	91.54	89.96	92.57
CART	91.00	85.38	94.65
ID3	89.67	87.36	91.17
MLP	93.15	89.74	95.37
SVM	89.97	82.23	95.01
RBF	82.80	89.90	78.18

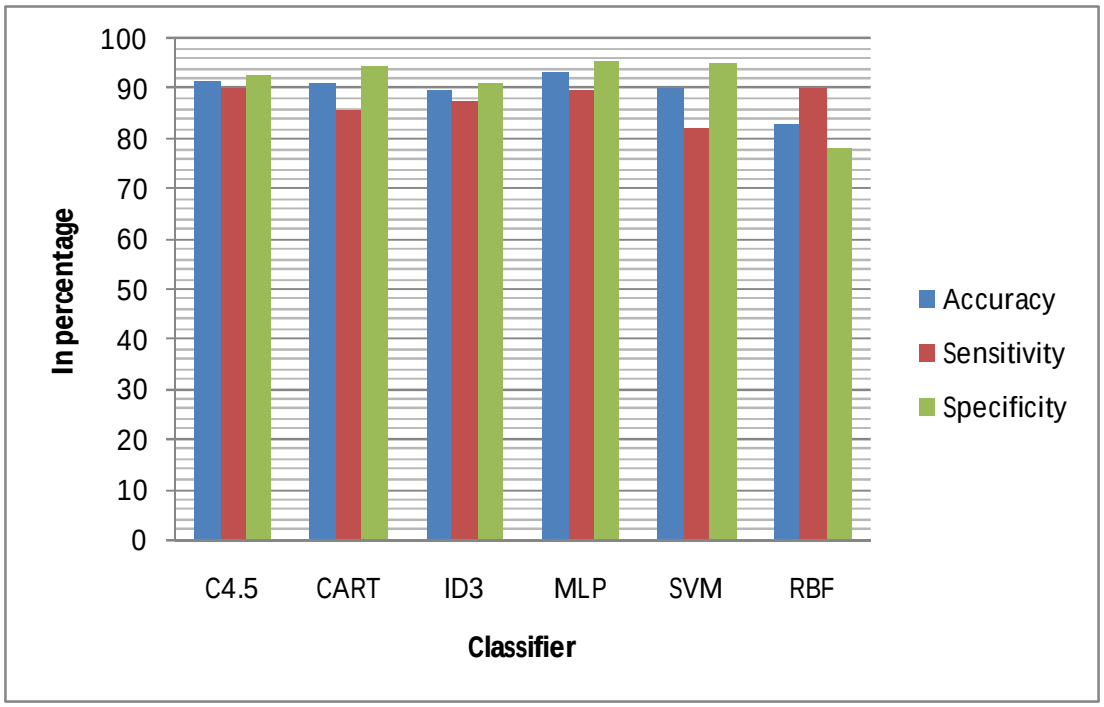


Fig 1. Performance measures of various models

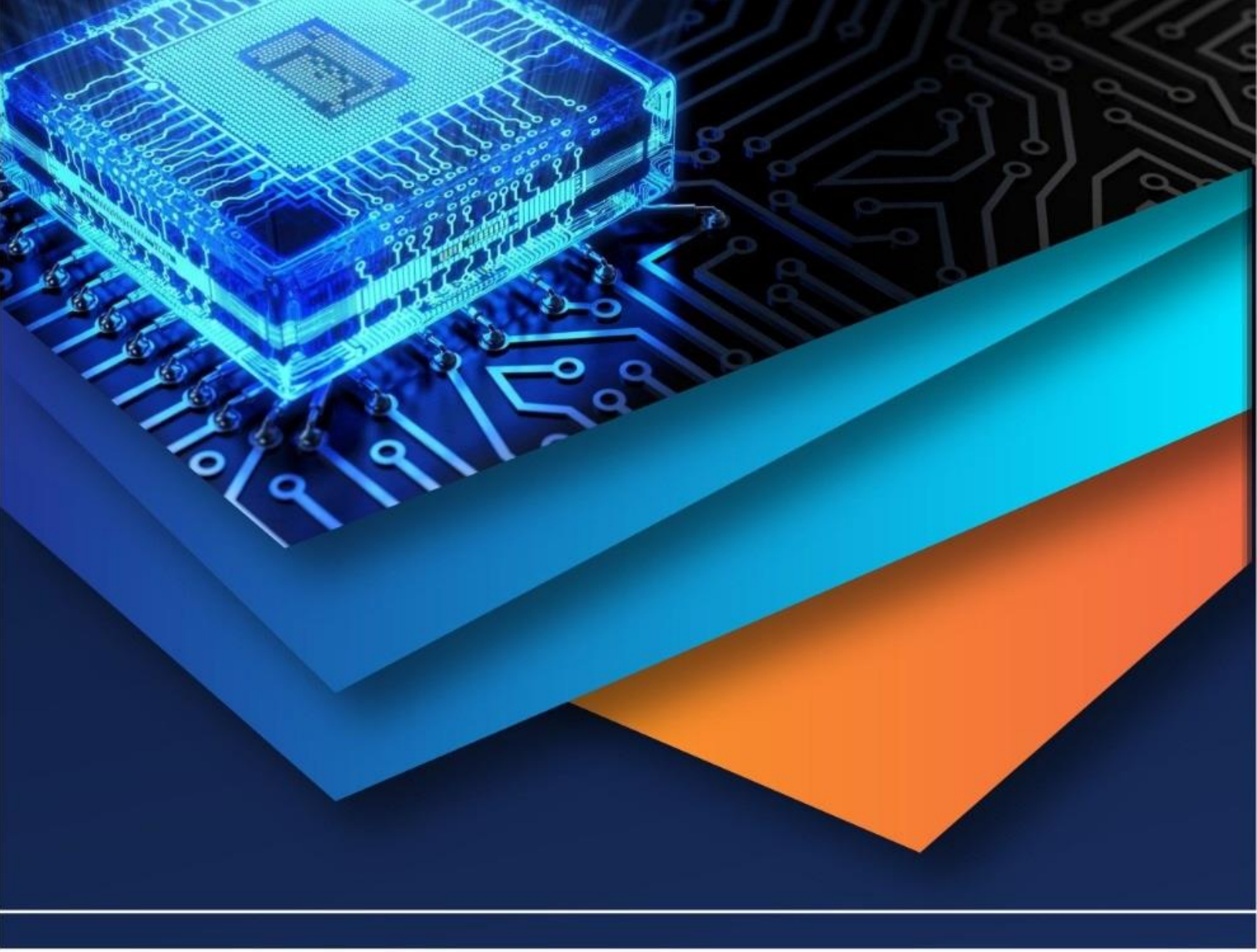
IV. CONCLUSIONS

E-mail is one of the important and easy communication media for sharing the information from one person to another person via the internet. In this research work, we have used data mining based classification techniques to classification of spam and non-spam data. We have recommended MLP as best classifier for classification of spam e-mail.

In future, we can develop hybrid model to improve the performance of models. We will also use various ranking based and others optimization techniques to computationally increase the performance of models.

REFERENCES

[1] Jiawei Han and Micheline Kamber,"Data Mining: Concept and techniques" Simon Fraser University,2005.
[2] A. K.Pujari A. K. Pujari, Data Mining Techniques. Universities Press (India) Private Limited. 4th ed., ISBN: 81-7371-380-4,2001.
[3] Y. Liu, Z. Zhenfang and Z. Jing, " A Word Sequence Kernels used in Spam-Filtering", Scientific Research and Essays, Vol. 6, pp. 1275-1280, 2011.
[4] L. Varghese, M. H. Supriya and J Poulose, "Filtering Templates Driven Spam mails using Vector Space Models", International Journal of Computer Applications, Vol. 39, pp. 33-35, 2012.
[5] M. Awad and M. Foqaha, " E-mail spam classification using hybrid approach of RBF Neural Network and Partical Swarm Optimization", International Journal of Network Security & Its Applications (IJNSA), Vol.8, No.4,2016.
[6] R. Sharma and G. Kaur, "E-mail Spam Detection Using SVM and RBF", International Journal of Modern Education and Computer Science, Vol. 4,pp. 57-63, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY