International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
August 2015, PP 15-19
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

Decision Tree Model for Classification of E-mail Data with Feature Selection

Akhilesh Kumar Shrivas
DLS PG College
Bilaspur
Chhattisgarh, India
akhilesh.mca29@gmail.com

Rahul Hota
M. Tech(CSE)
RGPV
Bhopal, India
rahulhota17@gmail.com

Abstract: *Due to frequent uses of internet, protecting of information from suspicious users are very challenging task for every internet users. Phishing and spam E-mail are one of the important and challenging issues faced by the world of e-commerce today. Phishing attacks are one of the emerging serious threats against personal data security. The aim of phishing is to steal a user's identity in order to make fraudulent transactions. Spam E-mail is junk E-mail send by spammers which unnecessary increase the communication bandwidth and wastage space on mail box. There are large number of techniques have been proposed and implemented for detecting phishing attacks and spam E-mail, but a complete solution is missing. In this experiment, we have used decision tree techniques for classification of phishing and spam E-mail. Both data sets applied on decision tree algorithm like CART, CHAID, QUEST and its ensemble models with three different partitions. Partitions of data set also play very important role for varying accuracy. Feature selection plays important role to increase the performance of model. An ensemble of CART and CHAID gives high accuracy as 99.29 % in case of phishing E-mail and 90.79% in case of spam E-mail data set which is higher than each individual model. The same model also gives same accuracy with 11 and 3 features in case of phishing and spam E-mail data set respectively. The proposed ensemble model is robust classifier for classification of phishing and spam E-mail.*

Keywords: Spam *E-mail, Ensemble Model, Feature Selection, Phishing, Information Gain.*

1. INTRODUCTION

Now days, due to increasing demand of internet, protection of information from unauthorized person are very challenging task for every internet users. Phishing and spam E-mail are very challenging issues for every internet users. A lots amount of research is being carried out to solve problem of phishing and spam E-mail and developed various tools but it is insufficient methods that can be used against phisher and spam novel attacks. Isredza Rahmi A Hamid et al. [1] have used various models like Bayesian Net, AdaBoost, Decision Tree and Random Forest for classification of phishing E-mail and Random forest gives 93% of accuracy. Almomani, A. et al. [2] and Yearwood , J. et al. [3] have also discussed the phishing mail classification. Shanmuga Priya , D. et al. [4] have used various classification techniques and feature selection techniques on spam E-mail data set to develop an efficient spam E-mail classifier. The BayesNet gives better performance than other techniques with accuracy of 86.7% in case of only 8 features.

An ensemble model and feature selection techniques play very important role to robustness of model. In this research work, we have used ensemble model and feature selection technique which given high classification accuracy as well as improve the performance of model.

2. TECHNIQUES

This research work used various techniques to classification of phishing and spam E-mail data as below:

2.1 Decision Tree

Decision tree [7] is the most popular data mining technique. The most common data mining task for a decision tree is classification .The principle idea of a decision tree is to split our data recursively into subsets so that each subset contains more or less homogeneous states of our target variable

(predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When this recursive process is completed, a decision tree is formed. There are following decision tree used in this research work:

CART [8] is one of the popular methods of building decision tree in the machine learning community. CART builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. CART uses the gini index for determining the best split.

QUEST [9] is a binary-split decision tree algorithm for classification and data mining. The objective of QUEST is similar to that of the CART algorithm. The major differences are: QUEST uses an unbiased variable selection technique by default. QUEST uses imputation instead of surrogate splits to deal with missing values. QUEST can easily handle categorical predictor variables with many categories.

CHAID [8] is decision tree algorithm proposed by Hartigan. CHAID attempts to stop growing the tree before overfitting occurs, and then carry out pruning as post processing step. In that sense, CHAID avoids the pruning phase.

2.2 Ensemble Model

An ensemble model combines the [5] output of several classifier produced by weak learner into a single composite classification. It can be used to reduce the error of any weak learning algorithm. The purpose of combining all these classifier together is to build a hybrid model which will improve classification accuracy as compared to each individual classifier. In this research work we have used voting scheme for ensemble models.

2.3 Feature Selection

Feature selection [6] is an optimization process in which one tries to find the best feature subset from the fixed set of the original features, according to a given processing goal and feature selection criteria. We have used Information gain [10] feature selection technique to select the relevant subset features from data set.

3. DATA SETS

In this research work, we have used phishing and spam E-mail data set collected from Spam assassin and UCI repository site respectively. The phishing data set contains 8266 instances, 48 features and 1 class having phishing and ham, similarly spam E-mail data set contains 4601 instances, 57 features and 1 class having spam and non-spam. There is no missing value in this data set.

4. PERFORMANCE MEASURES

The robustness of model can be check by various performance measures like sensitivity, specificity and accuracy. These measures are calculated using true positive(TP), true negative(TN), false positive(FP) and false negative(FN) which forms confusion matrix. The confusion matrix [10] for two classes is shown in Table 1.

Table 1. *Confusion matrix for positive and negative cases*

Actual Vs. Predicted	**Positive**	**Negative**
Positive	True Positiv (TP)	False Negative(FN)
Negative	False Positive (FP)	True Negative(TN)

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

$$\text{Sensitivity} = TP/(TP+FN) \quad (2)$$

$$\text{Specificity} = TN/(TN+FP) \quad (3)$$

5. RESULTS AND DISCUSSION

The experiment part is divided into two sections: first analysis of individual and ensemble models and second feature selection techniques applied on best model for classification of phishing and spam E-mail.

In this experiment, we have used decision tree classification techniques for classification of phishing and spam E-mail. Various classification techniques like CART, CHAID, QUEST and its ensemble models have applied on phishing and spam E-mail data set with different partitions for classification of phishing and spam E-mail shown in Table 2. An ensemble of CART and CHAID gives best 99.29% and 90.79% of accuracy in case of phishing and spam E-mail data set respectively. Figure 1 also show that bar chart which represent accuracy of model for phishing and spam E-mail with different data partitions.

Table 2. *Classification accuracy (%) of different models*

Techniques	75-25% partition		80-20% partition		90-10% partition	
	Phishing	Spam	Phishing	Spam	Phishing	Spam
CART	98.84	90.45	98.86	89.97	99.06	89.74
CHAID	98.65	88.66	98.07	89.01	97.05	87.99
QUEST	98.84	82.78	98.62	83.56	98.58	84.93
QUEST+CHAID	98.84	88.83	98.86	89.22	98.94	89.74
CART+ CHAID	98.89	**90.79**	98.68	90.39	**99.29**	90.39

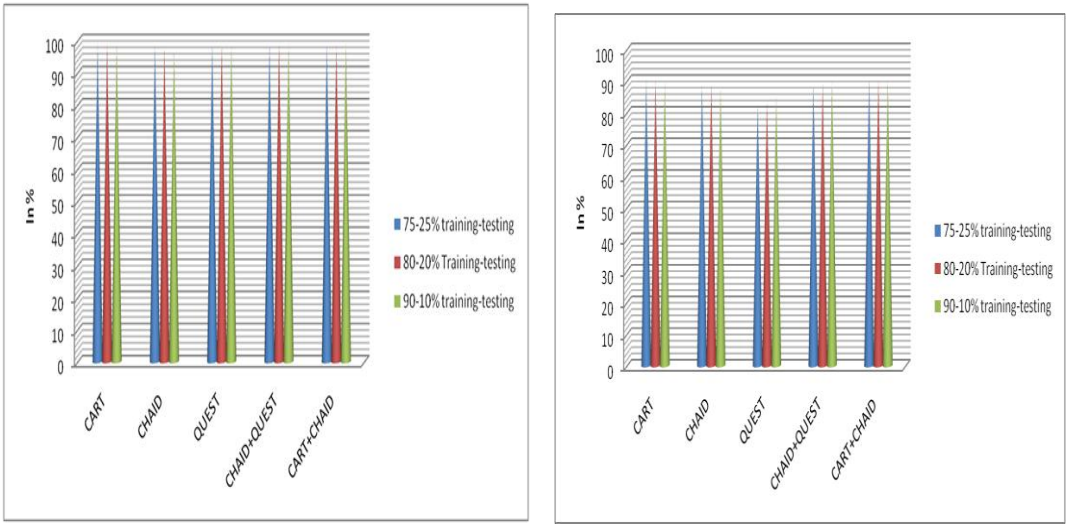

Figure1. *Accuracy of various models (a) Phishing E-mail (b) Spam E-mail*

To improve the performance of models, feature selection plays important role and achieved high accuracy with reduced number of features. In this experiment, we have used Information Gain feature selection technique for phishing and spam E-mail data classification. We have selected best model as ensemble of CART and CHAID for both phishing and spam E-mail classification. We have eliminated feature one by one from both data set and check the accuracy of models. With different feature subset of data set, accuracy of model is constant. Our proposed model gives same accuracy in case of reduced feature subset for both the data set. The proposed model gives 99.29 % of accuracy with 11 features in case of phishing E-mail data set, similarly model also gives 90.79% of accuracy with 3 features in case of spam E-mail data set shown in Table 3. Figure 2 show that accuracy of model with feature subsets in case of both data sets. Table 4 shows that confusion matrix of best model. Table 5 shows that various performance measures like accuracy, sensitivity and specificity are calculated with the help of confusion matrix using equation 1,2 and 3.Figure 3 also shows that various performance measures of best model. Finally our proposed model is better for classification of phishing and spam E-mail.

Table 3. *Information Gain feature selection technique*

Data Set	Number of features	Accuracy
Phishing E-mail	11	99.29
Spam E-mail	3	90.79

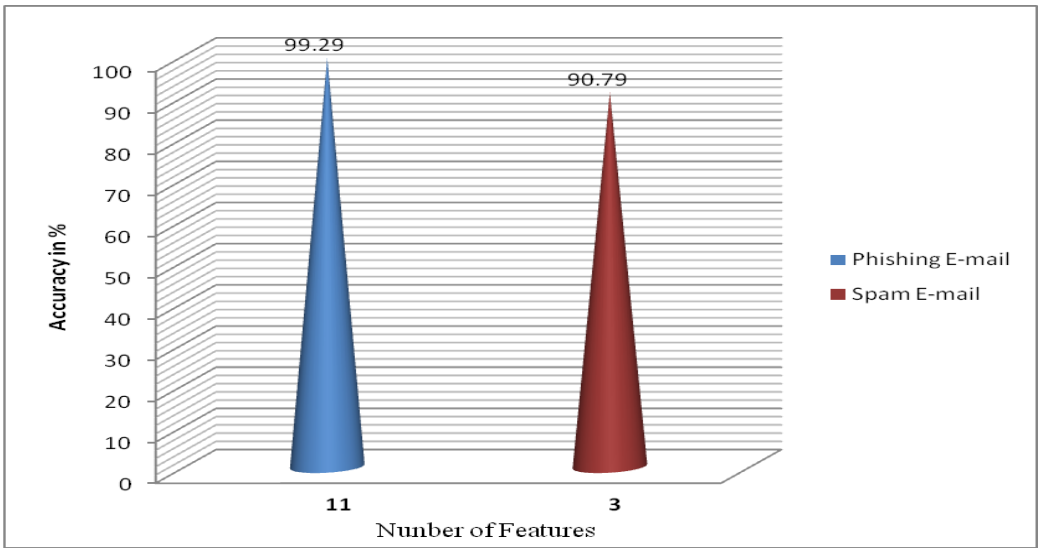

Figure 2. *Accuracy of best model with feature subsets*

Table 4. *Confusion matrix of best model*

Actual Vs Predicted	Phishing E-mail		Spam E-mail	
	Ham	Phishing	Non-Spam	Spam
Ham/Non-Spam	405	1	658	36
Phishing/Spam	5	437	72	407

Table 5. *Performance measures of the best model*

Performance Measures	Phishing E-mail	Spam E-mail
Accuracy	99.29%	90.79%
Sensitivity	99.75%	94.81%
Specificity	98.86%	84.96%

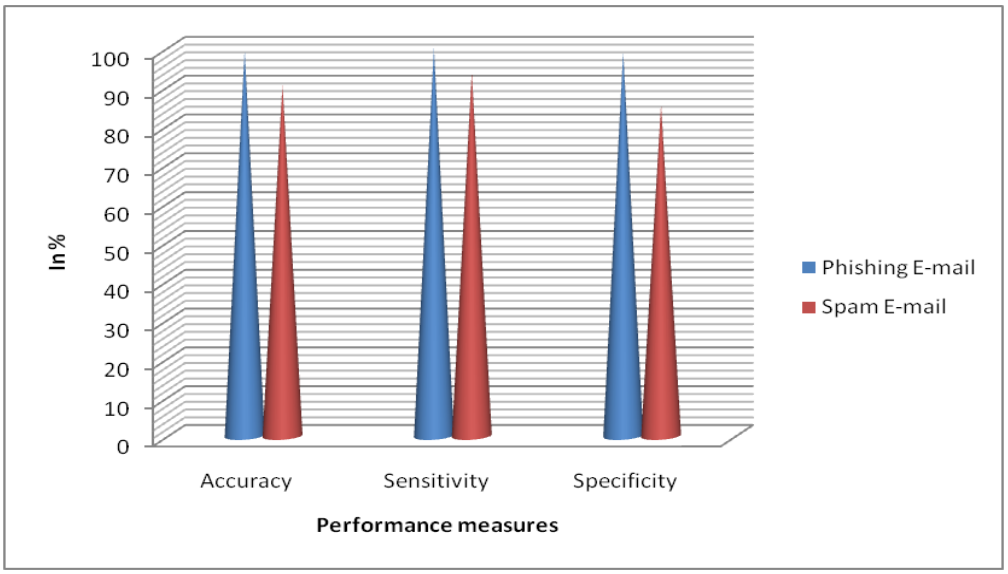

Figure 3. *Various performance measures of best model*

6. CONCLUSION

Information security is one the most important and challenging task for every Internet users. Every organization and industries are facing the problem of information security and need to secure information from unauthorized person. Classification of Phishing and spam E-mail are very challenging issues for every E-mail users. In this paper ,our proposed ensemble of CART and CHAID model gives 99.29% and 90.79% accuracy for phishing and spam E-mail classification respectively in case of Information gain feature selection technique with reduced feature subset. In future we can

apply genetic algorithm and partial swarm optimization technique to optimize problem and achieve high accuracy.

REFERENCES

[1]. Isredza Rahmi A Hamid and Abawajy J., Phishing Email Feature Selection Approach, 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11, DOI: 10.1109/TrustCom.2011.126, (2011).

[2]. Almomani, A., Wan, T., Ahmad, M. , Altaher , A., Almomani, E., Al-Saedi, K., Ahmad, A. and Ramadass ,S., A survey of Learning Based Techniques of Phishing Email Filtering, International Journal of Digital Content Technology and its Applications(JDCTA) ,6(18), (2012).

[3]. Yearwood, J., Mammadov, M. and Banerjee, A., Profiling Phishing E-mails Based on Hyperlink Information, 2010 International Conference on Advances in Social Networks Analysis and Mining, DOI: 10.1109/ASONAM.2010.56, (2010).

[4]. Shanmuga Priyaa , D., Kivitha, B., R. ,Naveen Kumar and Banuroopa, K., Improvising BayesNet Classifier Using Various Feature Reduction Method for Spam Classification, International Journal of Computer Science and Technology (IJCST), 1, (2010).

[5]. Pal, M. , Ensemble Learning with Decision Tree for Remote Sensing Classification, World Academy of Science, Engineering and Technology. 36, (2007).

[6]. Cios, K. J., Pedrycz ,W., and Swiniarski , R. W., Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 3rd ed., ISBN: 0-7923-8252-8, (1998).

[7]. Tang, Z., Maclennan, J., Data Mining with SQL Server 2005.Willey Publishing, Inc, USA, ISBN: 13: 978-0-471-46261-3, (2010).

[8]. Pujari, A. K., Data Mining Techniques. Universities Press (India) Private Limited. 4th ed., ISBN: 81-7371-380-4, (2001).

[9]. Yu-Shan, S. and Wei-Yin, L., QUEST Classification Tree (Version 1.9.2), http://www.stat.wisc.edu/~loh/quest.html . (Browsing date: 25th April 2012).

[10]. Han, J. and Kamber, M., Data Mining Concepts and Techniques. Morgan Kaufmann, San

[11]. Francisco. 2nd ed., ISBN: 13: 978-1-55860-901-3, (2006).