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

Decision Tree Classifier for Classification of Phishing Website with Info Gain Feature Selection

A. K. Shrivas [1], Ramkishun Suryawanshi[2]
[1,2] *Dept. Of IT, Dr. C. V. Raman Univrsity, Bilaspur (C. G.), India*

Abstract: ***Security of the information is very challenging task for every organizations and institute due to increasing demand of information and communication technology. Phishing attack is one of the important issues to access the sensitive information from unauthorized person. Data mining based classification intelligent techniques play very important role to classify the phishing and non phishing attack. In this research work, we have proposed decision tree technique and Info gain feature selection technique (FST) using different top selected feature subsets for developing computationally efficient model for classification of phishing websites. Our proposed Decision Tree (DT) technique gives better classification accuracy as 99.80% with 15 numbers of features in case of Info gain FST.***
Keywords: ***Classification, Phishing, Decision Tree, Info Gain.***

I. INTRODUCTION

Now days, security is one of the important issues due to increasing demand of internet and intranet. There are various internet and e-mail users are facing the problem of security of information. Phishing attacks are facing by every e-mail users which receive in mail with hyperlink which redirect to the phishing websites. Phishing websites harmful for users which collect the sensitive information from users like password, user id etc. Classification techniques play very important role to protect the information from unauthorized person,. Classification techniques are used to develop a classifier which classify the phishing and non phishing attacks. Some authors have worked in the field of classification of phishing websites. U. Naresh et al. [1] have studied about hyperlinks associated with phishing e-mails. They have proposed Link Guard algorithm to identify and classification of phishing e-mails associated with such type of hyperlink. R. Gupta et al. [2] have proposed new ePhishNet anti phishing tools and compared with others anti phishing tools. They have used Class Imbalance Problem (CIP), Rule based Classifier (Sequential Covering Algorithm (SCA)), Nearest Neighbour Classification (NNC) and Bayesian Classifier (BC) for analysis and identify the phishing attacks. J. Gori Mohamed et al. [3] have analyzed about phishing attacks and its disadvantages faced by the internet and e-mail users. They have discussed various phishing techniques and statistics of phishing in different years. V. Suganya [4] have discussed various types of phishing attacks that collect the sensitive information from users or harmful for users. They have also discussed various anti phishing techniques proposed by different researchers. M., Al-diabat (2016) [5] have suggested Decision tree (C4.5 and IREP) as classifier with Information gain and Symmetrical Uncertainty FST to develop robust and computationally increase the performance of model.

II. CLASSIFICATION TECHNIQUE

Classification is a supervised technique which is used to predict the class member of data set. In classification process, data set is divided into two sections as training and testing. Training data set is used to train the classifier and testing data set is used to test the classifier. In this research work, we have used Decision Tree (DT), Random Tree, Random Forest and Decision stump used as classifier for classification of phishing websites.

A decision tree (Source: help file of Rapid miner tool)[6] is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret. The goal is to create a classification model that predicts the value of a *target attribute* (often called *class* or *label*) based on several input attributes of the Example Set. Each interior node of tree corresponds to one of the input attributes. The number of edges of a nominal interior node is equal to the number of possible values of the corresponding input attribute. Outgoing edges of numerical attributes are labeled with disjoint ranges. Each leaf node represents a value of the *label* attribute given the values of the input attributes represented by the path from the root to the leaf.

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

Random Forest (or RF) (Parimala, R. et al., 2011) [7] is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random Forests are often used when we have very large training datasets and a very large number of input variables (hundreds or even thousands of input variables). A random forest model is typically made up of tens or hundreds of decision trees.

The Random Tree (Source: help file of Rapid miner tool) [6] works exactly like the Decision Tree with one exception: for each split only a random subset of attributes is available. This tree learns decision trees from both nominal and numerical data. Decision trees are powerful classification methods which can be easily understood. The Random Tree operator works similar to Quinlan's C4.5 or CART but it selects a random subset of attributes before it is applied. The size of the subset is specified by the *subset ratio* parameter.

The Decision Stump (Source: help file of Rapid miner tool) [6] is used for generating a decision tree with only one single split. The resulting tree can be used for classifying unseen examples. This operator can be very efficient when boosted with operators like the AdaBoost operator.

III.DATA SET

We have used phishing website data set that is collected from UCI repository [8]. The data set contains 30 features, 11055 instances and 1 class having phishing website and non phishing websites. In this research work, we have identified feature name with feature id from 1 to 30.

IV.RESULTS AND DISCUSSION

In this experiment, we have used Rapid miner data mining tool in window 7 operating system environment with i3 processor. We have used data mining based classification techniques like Decision Tree(DT), Random Tree , Random Forest and Decision Stump for classification of phishing and non phishing websites with 10-fold cross validation. We have also compared the accuracy of models and achieved better classification accuracy with Decision Tree classifier. Table I show that accuracy of various models where Decision Tree achieved best accuracy as 91.80%. Fig. 1 shows that accuracy of models in the form of bar chart.

Dimensional reduction is very important role to computationally increase the performance of model. Feature selection is used to reduce the dimension using remove the feature from original feature space. In this research work have used Info gain FST to select top relevant feature subset and applied on best Decision Tree classifier. Table II shows that selected top feature subset 5, 10, 15 and 20 with 91.70%, 91.75%, 91.80% and 91.80% accuracy respectively with Decision Tree classifier. Confusion matrix of selected top feature subset as shown in table III with Decision Tree classifier. Various performance measures like sensitivity, specificity and accuracy are calculated with different feature subset as shown in table IV. We have achieved satisfactory results as 91.80% of accuracy, 91.18% of specificity and 92.27% of specificity with 15 feature subset.

Table I
Accuracy of models with 10-fold cross validation

Model	Accuracy
Decision Tree (DT)	91.80
Random Tree	66.75
Random Forest	78.85
Decision Stump	84.73

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

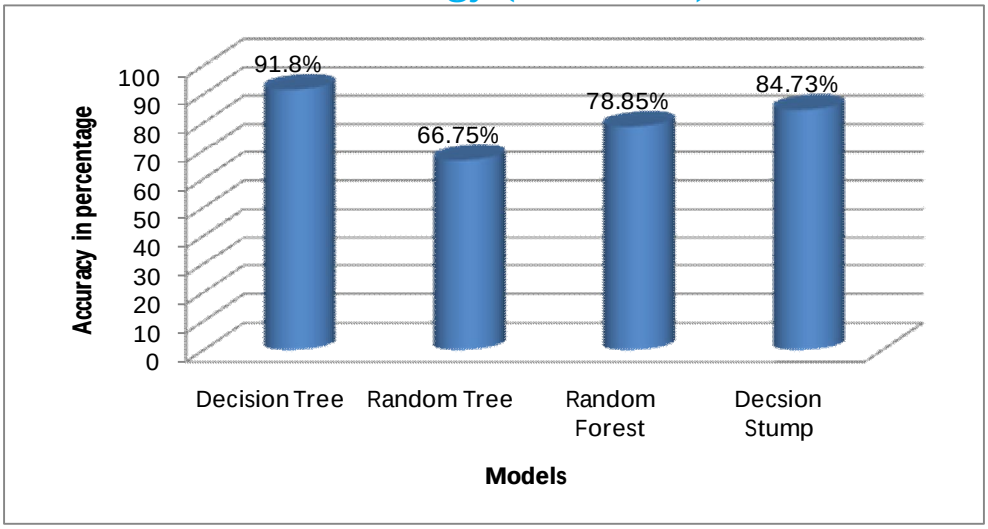

Fig. 1 Accuracy of various models

Table II
Selected top features subset using Info Gain FST

Number of features	Feature subsets with feature_id	Accuracy
5	6,7,8,14,26	91.70
10	6,7,8,9,13,14,15,16,26,28	91.75
15	1,3,6,7,8,9,13,14,15,16,24,26,27,28,30	91.80
20	1,2,3,4,6,7,8,9,13,14,15,16,18,24,25,26,27,28,29,30	91.80

Table III
Confusion matrix of selected top feature subset with Info gain FST

Actual Vs. Predicted	5 feature subset		10 feature subset	
	NPW	PW	NPW	PW
NPW	4423	443	4416	430
PW	475	5714	482	5727
Actual Vs. Predicted	15 feature subset		20 feature subset	
	NPW	PW	NPW	PW
NPW	4418	427	4421	429
PW	480	5730	477	5728

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

Table IV
Performance measures of selected top feature subset with Info gain FST

Number of features	Accuracy	Sensitivity	Specificity
5	91.70	90.89	92.32
10	91.75	91.12	92.23
15	91.80	91.18	92.27
20	91.80	91.15	92.31

V. CONCLUSIONS

Phishing attack is very serious problem for internet users and face by e-mail users. Classification is important technique is used to identify and classification of phishing and non phishing attacks. In this research work have used data mining based classification techniques for classifying phishing and non phishing attacks. We have also applied Info gain feature selection technique to reduce the feature subset and computationally increase the performance of model. We have recommended Decision Tree classifier for classification of phishing attacks with few numbers of features.

REFERENCES

[1] U. Naresh, U.Vidya Sagar and C. V. M. Reddy , "Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm", IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 14, Issue 3 , pp. 28-36, 2013.
[2] R. Gupta and P. K. Shukla, "Performance Analysis of Anti-Phishing Tools and Study of Classification Data Mining Algorithms for a Novel Anti-Phishing System", International journal of Computer Network and Information Security, Vol. 12, pp. 70-77,2015.
[3] J. Gori Mohamed, M. Mohammed Mohideen and .N. Shahira Banu , "E-Mail Phishing - An open threat to everyone", International Journal of Scientific and Research Publications, Vol. 4, Issue 2, pp. 1-4, 2014.
[4] V. Suganya, "A Review on Phishing Attacks and Various Anti Phishing Techniques", International Journal of Computer Applications, Vol. 139 ,No.1,pp. 20-23, 2016.
[5] M. Al-diabat , "Detection and Prediction of Phishing Websites using Classification Mining Techniques", International Journal of Computer Applications, Vol. 147, No.5, pp. 5-11, 2016.
[6] Source: Help Fife: Rapid mining data mining tools (Browsing date 20-04-2017)
[7] R. Parimala, and R. Nallaswamy, "A Study of Spam e-mail Classification using Feature Selection Package", Global Journal of Computer Science and Technology", Vol. 11, Issue 7, 2011.
[8] UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html].(Browing date: 20-03-2017).