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

DOI: http://doi.org/10.22214/ijraset.2017.4205

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

# An Ensemble Model for Identification of Phishing Website

Jaspreet Kaur Virdi[1], Amit Kumar Dewangan[2]
*Dept. of CSE, Dr. C. V. Raman University, Kota, Bilaspur, Chhattisgarh, India*

***Abstract: Identification of phishing websites is very challenging task for every internet and e-mail users. To protect the information from unauthorized person classification of phishing websites is very important. In this research work , we have used many data mining based classification techniques like C4.5, SimpleCart, Random tree, SVM and MLP for classification of phishing websites with different data partitions like 75% training and 25% testing,, 80% training and 20% testing and 85% training and 15% testing. To develop a robust model , we have ensemble the models with different combinations. We have achieved better accuracy with ensemble of C4.5, SimpleCart , MLP and Random tree with all data partitions, but it achieved best accuracy as 97.16% in case of 85-15% data partition.***

***Keywords: Ensemble model, Classification, Phishing Websites.***

## I. INTRODUCTION

Today’s, increasing number of internet and e-mail users, security of information is very import issues. Phishing websites originates from phishing e-mails that contain the suspicious link which collect the sensitive information of authorized users by the unauthorized person. To protect the information from unauthorized person, classification of phishing websites is very challenging task. There are various authors have worked in the field of classification of phishing websites. K. Rajitha et al. (2016) [5] have analyzed the malicious detection problems .They have offered a survey of the malicious website detection techniques using various phishing method. M. Al-diabat et al. (2016) [6] have investigated features selection aiming to determine the effective set of features in terms of classification performance. They compare two known features selection method in order to determine the least set of features of phishing detection using data mining. S. Khairnar et al. (2016) [7] have investigated different Online Fraud Transaction prevention system is studied based visual cryptography. From study we proposed a method for Online Fraud Transaction prevention using EVC and QR code techniques. R. Islam et al. (2013) [8] proposed a new approach called multi-tier classification model for phishing email filtering. They also propose an innovative method for extracting the features of phishing email based on weighting of message content and message header and select the features according to priority ranking. They will also examine the impact of rescheduling the classifier algorithms in a multi-tier classification process to find out the optimum scheduling The results of the experiments show that the proposed algorithm reduces the false positive problems substantially with lower complexity. Suganya (2016) [9] discussed about the various types of phishing attacks and various anti phishing techniques used to prevent phishing attack. A. K. Shrivas et al. (2015) [10] have used various decision tree based classification techniques and its ensemble model for classification of spam and phishing e-mail data. The proposed ensemble of CART and CHAID give better accuracy with both spam and phishing e-mail data set.

## II. METHODS AND MATERIALS

This section includes various data mining based classification techniques for classification of data. They have also described the phishing website data set used in this research work.

### *A. Decision Tree*

Decision tree is data mining based classification techniques to generate the rules and classification of data based on this rule. Decision tree (*Han J.et al.., 2006*) **[2]** 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 research work, we have used C4.5, SimpleCart and Random tree as decision tree.

### *B. Support Vector Machine (SVM)*

A SVM (Vapnik, V. , 1998) [3] is a promising new method for classification of both linear and nonlinear data. SVM is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM algorithms divide the n dimensional space representation of the data into two regions using a hyper plane. This hyper plane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two classes and is computed based on the distance between the closest instances of both classes to the margin, which are called supporting vectors.

### *C. Bayesian Net*

Bayesian classifiers (Han, J. et al., 2006) [2] are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Classification algorithms have found a simple Bayesian classifier known as the Naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

### *D. Multilayer Perceptron (MLP)*

MLP (Pujari, A. K., 2001) [1] is a development from the simple perceptron in which extra hidden layers (layers additional to the input and output layers, not connected externally) are added. More than on hidden layer can be used. The network topology is constrained to be feed forward, i.e., loop-free. Generally, connections are allowed from the input layer to the first (and possible only) hidden layer, from the first hidden layer to the second and so on, until the last hidden layer to the output layer. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The actual construction of network, as well as the determination of the number of hidden layers and determination of the overall number of units, is sometimes of a trial-and-error process, determined by the nature of the problem at hand. The transfer function generally a sigmoid function.

### *E. Ensemble Technique*

Two or more modes combined to form a new model is called an ensemble model (Han J. et al., 2006) [2] . An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. In this research work, we have used various combination of C4.5, SimpleCart and Random tree, SVM,MLP and bayes net to improve the performance of model.

### *F. Data Set*

In this research work, we have used phishing website data set collected from UCI repository [4] . The data set consist 30 features and 1 class having phishing website and non -phishing websites. The data set consist 11055 records.

## III. EXPERIMENT RESULTS

In this experiment, we have used various classification techniques like C4.5, SimpleCART, Random Tree, SVM and MLP as classifier for classification of phishing websites. Partitions of data is also one of the important role for varying accuracy. The accuracy of Random tree gives better accuracy in case all the partition as shown in table 1. To achieve the better classification accuracy, we have ensemble the two or more models with 75-25% training-testing partition, 80-20% training-testing partition and 85-15% training-testing partition. The accuracy of ensemble models with different partitions as shown in table 1. All the ensemble models give better accuracy with all selected partition, but we have achieved best accuracy in case of ensemble of C4.5, SimpeCART , MLP and Random tree. We have achieved the best accuracy with proposed ensemble of C4.5, SimpeCART , MLP and Random tree as 96.16%. Table 2 shows that confusion matrix of best model with all partitions. We have also calculated the various performance measures like sensitivity and specificity of best model with the help of confusion matrix as shown in table 3. The sensitivity is highest in case of 75-25% training-testing partition while specificity is highest with 85-15% training-testing partition. Finally, we can recommended, our proposed model gives better classification accuracy for classifying the phishing websites.

Table 1: Accuracy Of Models With

| Model | 75-25% | 80-20% | 85-15% |
|---|---|---|---|
| C4.5 | 96.3459 | 95.9294 | 95.7177 |
| SimpleCart | 95.6223 | 95.7033 | 95.778 |
| Random tree | 96.4544 | 96.427 | 96.3209 |
| SVM | 92.2576 | 92.3112 | 92.3402 |
| MLP | 95.9841 | 96.1556 | 96.2606 |
| Bayesnet | 92.7641 | 92.6278 | 92.4608 |
| C4.5 + SimpleCart | 96.5268 | 95.9747 | 95.959 |
| MLP + SimpleCart | 96.4544 | 96.5174 | 96.924 |
| C4.5 + Random tree | 96.8524 | 96.6079 | 96.5621 |
| C4.5 +Simple CART + Random tree | 96.78 | 96.3817 | 96.8034 |
| C4.5 +SimpeCART + MLP+Random tree | 97.1056 | 96.7436 | 97.1653 |

Table 2: Confusion matrix of best ensemble model (C4.5 +SimpleCART + MLP+Random tree)

| Actual Vs. Predicted | 75-25% data partition | | 80-20% data partition | | 85-15% data partition | |
|---|---|---|---|---|---|---|
| | Non Phishing | Phishing | Non Phishing | Phishing | Non Phishing | Phishing |
| Non Phishing | 1179 | 56 | 946 | 55 | 706 | 39 |
| Phishing | 24 | 1505 | 17 | 1193 | 8 | 905 |

Table 3: Performance measures of best ensemble model (C4.5 +SimpeCART + MLP+Random tree)

| Performance measures | 75-25% data partition | 80-20% data partition | 85-15% data partition |
|---|---|---|---|
| Accuracy | 97.10 | 96.74 | 97.16 |
| Sensitivity | 95.46 | 94.50 | 94.76 |
| Specificity | 98.43 | 98.76 | 99.12 |

## IV. CONCLUSION

Protecting the information or information system from unauthorized is very challenging task. A Phishing attack is very critical problem faced by every e-mail users. Classification is one of the important issues to classify the phishing and non-phishing attacks. Partition of data is one of the important role for classification accuracy. In this research work, proposed ensemble (C4.5+SimpleCart+MLP+Random tree) is robust and efficient model and recommended for classification of phishing websites with 80-15% training-testing data partition.

## REFERENCES

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

[2] J. Han and M. Kamber , Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco. 2nd ed., ISBN: 13: 978-1-55860-901-3, 2006.

[3] V. Vapnik , Statistical Learning Theory, Wiley, 1998.

[4] UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science. Available: http://www.ics.uci.edu/~mlearn/databases/ (Browsing date: 15 Jan 2017).

[5] K. Rajitha and D. VijayaLakshmi , Comprehensive Study And Analysis Of Malicious Website Detection Techniques, International Journal of Computer Application, **Vol.** 6 No.5, pp. 7-21, 2016.

[6] M. Al-diabat , Detection and Prediction of Phishing Websites using Classification Mining Techniques, International Journal of Computer Application, Vol. 147, No. 5, pp. : 5-11,2016.

[7] S. Khairnar, Anti-Phishing framework based on Extended Visual Cryptography and QR code , International Journal of Computer Application, Vol. 142, No. 5, pp. 25-28,2016.

[8] R. Islam and J. Abawajy ,A multi-tier phishing detection and filtering approach, Journal of Network and Computer Applications. Vol. 36, pp. 324-335, 2013.

[9] V. Suganya ,A Review on Phishing Attacks and Various Anti Phishing Techniques, International Journal of Computer Applications, 139 (1) : 20-23,2016.

[10] A. K. Shrivas and R. Hota, Decision Tree Model for Classification of E-mail Data with Feature Selection, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), pp. 15-19, ISSN 2349-4859 (Online), 2015.