Intelligent Phishing Website Detection using Random Forest Classifier

Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery
Effat University, College of Engineering
Jeddah, 21478, Saudi Arabia
E-mail: {absubasi; emolah; fmalmakallawi, tchaudhery} @effatuniversity.edu.sa

***Abstract*-Phishing is defined as mimicking a creditable company's website aiming to take private information of a user. In order to eliminate phishing, different solutions proposed. However, only one single magic bullet cannot eliminate this threat completely. Data mining is a promising technique used to detect phishing attacks. In this paper, an intelligent system to detect phishing attacks is presented. We used different data mining techniques to decide categories of websites: legitimate or phishing. Different classifiers were used in order to construct accurate intelligent system for phishing website detection. Classification accuracy, area under receiver operating characteristic (ROC) curves (AUC) and F-measure is used to evaluate the performance of the data mining techniques. Results showed that Random Forest has outperformed best among the classification methods by achieving the highest accuracy 97.36%. Random forest runtimes are quite fast, and it can deal with different websites for phishing detection.**

Keywords- Web threat; Phishing Website; Random Forest Classifier; Data Mining Techniques.

I. INTRODUCTION

Different types of web threats may yield identity theft, stealing of private information, financial loss and loss of customer's assurance in online banking and e-commerce. Hence, internet suitability for commercial transactions becomes hesitant. Phishing is a form of web threats which is defined as the art of impersonating a website of a creditable company. Fraudulent persons create phishing websites to mimic web pages of genuine websites. The honest internet users are defrauded by these phishing websites which are similar to the legitimate ones. Since the number of phishing websites are increasing day by day, it will become a serious problem, even for the experienced users in the computer security and internet [1].

Detecting this type of attacks is a crucial step for online trading. Since many users think that they are safe against phishing attacks if an anti-phishing tool is used. Hence there is a great concern on the anti-phishing tools to detect phishing precisely. Phishing attack typically starts by sending an email which appears to come from a creditable company asking from the victims to confirm or update their private information by visiting the link in the e-mail. Different set of ways are used to produce phishing websites by the phishers, even if they are now using numerous techniques in forming phishing websites. In order to distinguish honest and phishing websites, two different approaches are used in recognizing phishing websites. The first one checks if the requested URL is on the blacklists by comparing with those in that list [2]. Meta-heuristic methods in which quite a lot features are collected from the website to categorize it as either legitimate or phishing website is the second approach [3]. The accurateness of the meta-heuristic method is based on extracting a set of distinguishing features which may help in differentiating the website [4]. Data mining techniques are generally used to extract the features from the websites to find patterns as well as relationships between them [5]. Data mining algorithms are highly imperative for decision-making, since decisions can be made based on the rules accomplished from a data-mining algorithm, [6]. In this paper we compared different data mining algorithms to detect the phishing websites and compare their performances by using classification accuracy, area under the ROC curve and F-measure.

This paper is structured as follows: Section II discusses works and different methods presented in the literature for phishing detection. Section III introduces used phishing data features, and methods. Finally, several experiments are described to measure the significance of the proposed classification algorithms in detecting phishing websites in Section IV. Conclusion is given in Section V.

II. RELATED WORKS

An intelligent system for phishing webpage detection in e-banking is proposed by Aburrous et al. [7]. They proposed a model based on fuzzy logic combined with data mining algorithms to examine the techniques by describing the phishing website aspects and by categorizing the phishing types. By using 10-fold cross-validation, they achieved 86.38% classification accuracy, which is very low. He et al. [8], proposed a model based on HTTP transaction, page content, and search engine results, they detected phishing pages with 97% of classification accuracy. A new type of intelligent algorithm based on approximate string matching is used by Arade et al. [9] to compare the addresses in the database of the proposed system and the webpage address. In this study, the problem is with the probability of occurring false positive occurrence, means legitimate webpages can be considered as phishing webpages. A model for detecting phishing webpages is proposed by Shahriar & Zulkernine [10] using

978-1-5386-0873-2/17/$31.00 ©2017 IEEE

the reliability of suspected pages. In their study, a finite state machine is proposed to assess webpage behavior by tracing the webpage form the submission as well as from the corresponding responses. MCAR is presented as a phishing detection method by Ajlouni et al. [11] by adopting the features from Aburrous et al. work by achieving 98.5% accuracy in classifying the webpages, but they did not give any information about how many rules were extracted by using the MCAR algorithm. A rule-based model in which Neuro-Fuzzy classifier with five inputs employed to detect phishing websites was proposed by Barraclough et al. [12]. The proposed model accuracy was 98.5%. Another approach, which uses the webpage under scrutiny and distinguish all the direct and indirect links related to the page, was proposed by Ramesh et al. [13]. The indirect page links are taken out from the search engine result but the direct links are taken out from the page content itself. In order to map the domains of suspicious webpage and phishing target related to IP, third-party DNS lookup is also used. They achieved 99.62% accuracy to detect phishing webpages but the method has external dependency which is 3rd-party DNS lookup and search engine result. Moreover, phishing webpages hosted on the compromised domains cannot be detected. Another detection model with a set of conventional features is proposed by Mohammed et al. [6] and calculated the detection error-rate yielded by the set of associative classification algorithms. The results presented that C4.5 has an average error-rate of 5.76%. In order to extract the rules from training data, Abdelhamid et al. [14] proposed a Multi-label Classifier based Associative Classification (MCAC). The limitation of the proposed model is that the induction of rules needs a large number of rules. They achieved 97.5% classification accuracy. Zhang et al. [15] used Sequential Minimal Optimization classifier with five features to distinguish Chinese phishing websites. The limitation of this approach is that the extracted features are only for detection of phishing webpages with Chinese language [16]. Li et. al. [17] used transductive support vector machine to detect and classify phishing web pages. They extract the features of web page image to reflect the characteristics of web pages absolutely. Montazer et. al. [18] used fuzzy logic combined with rough sets-based data mining algorithm for phishing detection. A method based on the differences between the phishing websites and the imitated target websites was proposed Li et. al. [19] and they used the ball-based SVM algorithm to distinguish phishing website. Moghimi et al. [16] used approximate string matching algorithms with all individual page resource elements and page hyperlinks instead of comparing them directly.

III. MATERIALS AND METHODS

A. Phishing Websites Data

In this article, we used the publicly available phishing websites data set from the UCI machine learning repository [20]. The data is prepared and donated by Mohammad et. al. [21] [22] [23] [1]. The features are used in the dataset is described in [22]

B. Artificial Neural Networks (ANN)

Artificial neural networks (ANN), inspired from biological neural networks, is a set of interconnected nodes (neurons). Each connection between nodes is typically assigned weights. The network learns by adjusting the weights, in the learning phase for correct prediction process. The long training times of ANNs makes them useful for applications that involve pattern recognition, clustering and classifying. Initially ANNs were considered less suitable for data mining due to their poor interpretability and long training times. However, their advantages include ability to classify patterns on which they have not been trained and high tolerance for noisy data. In addition to that, due to their parallel nature, parallelization techniques can speed up the computational process and rule extraction can be done through certain techniques. Hence, these factors play an important part in classification and numeric prediction in data mining [5].

C. K-Nearest Neighbour (k-NN)

Learning for k-NN classifiers occurs by analogy, that is, by comparing the test tuple to similar training tuples. These are distance-based comparisons that intrinsically assign equal weights to each attribute; therefore, accuracy could be poor when noisy or irrelevant data is presented. However, methods of editing and pruning have been introduced to solve the problem of useless and noisy data tuples respectively. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. The good value for the number of neighbors can be determined experimentally. Firstly, the error rate of the classifier is checked for every neighborhood value and incremented every time to allow for one more neighbor. The value that gives the minimum error rate can be selected [5].

D. Support Vector Machine (SVM)

Support vector machines (SVMs) is used for the classification of both linear and nonlinear data. In short, when given an original training data, the algorithm uses a nonlinear mapping to transform it into a higher dimension. In this dimension, a linear optimal hyperplane is searched, to keep the data of any two classes separate. SVMs can be used for classification and numeric prediction as well. The simplest form of SVM is a two-class problem, where the classes are linearly separable. For a 2-D problem, a straight line can be drawn to separate the classes, in fact, multiple lines could be drawn [5]. However, finding a perfect class for millions of training data set is time consuming and various tuning parameters, including kernel is used. Therefore, more research is required for finding the best kernel for a given data set and dealing with multiclass cases efficiently [5].

E. C4.5 Decision Tree

A decision tree is a directed, acyclic graph with two types of nodes, namely; internal nodes that represents a test and terminal node holds a class label. The branches represent the outcome of the test. The topmost node is the root node. Few of the advantages of decision trees are: they are easier to interpret; they require no domain knowledge, they are able to handle both numerical and categorical data as well as multi-output problems. In addition, they can easily be converted to classification rules and perform simple and fast learning and classification steps. Decision tree induction is a process of learning of the decision tree by recursive splitting of the dataset from the root onwards. When a data tree is formed, there are certain anomalies in it due to noise and outliers, therefore *tree pruning* is performed to remove them. The pruned trees are smaller and less complex in return. There are two types of approaches to tree pruning. First is *pre-pruning*, where any node is halted from splitting further. So, the internal node becomes the terminal node. Second one is *post-pruning*, which removes a subtree from a fully-grown tree [24].

F. Random Forests (RF)

Leo Breiman first introduced Random forests in his paper, as a method of building forest of uncorrelated trees using CART like method [25]. Random Forests can be built in tandem with random attribute selection using bagging. Random Forests follow an ensemble approach to learning, that is a divide and conquer approach for improving performance. In a simple decision tree, the input or test in added at the top and it traverses down the tree, ending up in smaller subsets. In a random forest, the ensemble mechanism combines various random subsets of trees. The input/test traverses through all the trees. The result is calculated based on average or weighted average of the individual results, or the voting majority in case of categorical data. The accuracy of a random forest depends on a measure of the dependence between the classifier and the strength of the individual classifiers and they improve the problem of overfitting of the decision trees [5].

G. Rotation Forest (RoF)

Rotation Forest (RoF) is an ensemble classifier in which the training data is created by randomly splitting the feature set into K subsets and Principal Component Analysis (PCA) is applied to each subset for any base classifier. RoF, which is derived from the Random Forest idea but is based on feature extraction, is considered to be more accurate than bagging. It trains each tree for the whole data set in a rotated feature space. Its main purpose is to build accurate and diverse classifiers. For the training of the individual classifiers, Bootstrap samples are taken. Then feature extraction is applied and thus a full feature set for each classifier in the ensemble is constructed [26]. The concept of the rotation approach is used to encourage simultaneously diversity and individual accuracy within the ensemble. All classifiers can be trained in parallel, which is also the case with Bagging and Random Forests [27].

IV. RESULTS AND DISCUSSION

In this study, we used publicly available Phishing websites data sets from the UCI machine learning repository[1] to assess the model performances and, we used open source WEKA[2] machine learning tool to test the accuracies, F-measure and area under the ROC curve of each method. Different data mining techniques are used for classification with a 10-fold cross validation.

In this study, accuracy is used to check the performances of the classifiers:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} x100\% \quad (1)$$

where TP, FP, TN and FN are number of true positives, false positives, true negatives and false negatives respectively. Besides area under the ROC (Receiver Operating Characteristic) curve (AUC) is used to evaluate the performance of the classifiers [5]. The success of classifier is depends on the AUC value which provides a sign of a characteristic values created by means of the quantified input data, and demonstrate how result is predicted reliably [28] [29] [30]. F-measure is also another performance measure and defined as follows:

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

where $$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \quad (3)$$

Different machine learning tools namely, Artificial Neural Networks (ANN), k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), C4.5 Decision Tree, Random Forest (RF), and Rotation Forest (RoF) are used as a classifier for detection of phishing website and results are presented in Table 1. Random forests gives the superior classification accuracy with 97.36%, k-NN is the second one with 97.18 %. SVM gives 97.17%, ANN gives 96.91%, Rotation Forest gives 96.79%, C4.5 gives 95.88%, CART gives 95.79%, and NB is the last one with 92.98% accuracy.

Another performance evaluation method is F-measure, which is coincident with our classification results. It is apparently seen from Table 1 that F-measure and the total accuracy results are close to each other for random forest classifier which approves the consistency of results. So, the obtained F-measure result is 0.974 which is coincident with total accuracy (97.36 %). This is the situation for all other machine learning algorithms as well.

The performance of the classifiers can also be assessed by the AUC. AUC of RF is the best (0.996) and followed by ANN (0.995) then rotation forest

[1] https://archive.ics.uci.edu/ml/datasets/Phishing+Websites#

[2] http://www.cs.waikato.ac.nz/ml/weka/

(0.994), k-NN (0.989), C4.5 (0.984), CART and NB (0.981) and SVM is the last (0.97).

The classifiers' performance in terms of F-measure, accuracy, and AUC showed that the machine learning algorithms were successful in detecting phishing attacks. The results are consistent for the all machine learning algorithms, and designated that a RF model can produce reliable results with high accuracy. Our study presented that machine learning tools could be used to detect phishing attacks effectively.

V. Conclusion

Phishing is a way to deceive via fake e-mails and websites to steal people's private information. Phishing prevents individuals from carrying their activities via the internet. Phishing website detection is crucial for the internet community since it has big impression on online transactions performed [14]. RF is an intelligent machine learning method that recently payed attention from researchers due to its speed and high classification accuracy. The phishing website problem has been investigated in this study in which we developed a machine learning model to determine correlations between the features and yields them from simple and effective rules. In this study, we adopted classifier model that is used for detecting phishing websites in an intelligent and automated way by using publicly available dataset. The performance of the proposed RF classifier is rather high in terms of classification accuracy, F-measure and AUC. Furthermore, our results showed that RF is faster, robust and more accurate than the other classifiers. Random forest's runtime is quite fast, and it is able to detect phishing website in comparison to the other classifiers.

References

[1] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, 2014.

[2] S. Gastellier-Prevost, G. G. Granadillo, and M. Laurent, "Decisive heuristics to differentiate legitimate from phishing sites," presented at the Network and Information Systems Security (SAR-SSI), 2011 Conference on, 2011, pp. 1–9.

[3] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," *ACM Trans. Inf. Syst. Secur. TISSEC*, vol. 14, no. 2, p. 21, 2011.

[4] N. Sanglerdsinlapachai and A. Rungsawang, "Using domain top-page similarity feature in machine learning-based web phishing detection," presented at the Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on, 2010, pp. 187–190.

[5] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[6] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Intelligent rule-based phishing websites classification," *IET Inf. Secur.*, vol. 8, no. 3, pp. 153–160, 2014.

[7] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7913–7921, 2010.

[8] M. He *et al.*, "An efficient phishing webpage detector," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12018–12027, Sep. 2011.

[9] M. S. Arade, P. Bhaskar, and R. Kamat, "Antiphishing model with url & image based webpage matching," *Int. J. Comput. Sci. Technol. IJCST*, vol. 2, no. 2, pp. 282–286, 2011.

[10] H. Shahriar and M. Zulkernine, "Trustworthiness testing of phishing websites: A behavior model-based approach," *Spec. Sect. SS Trust. Softw. Behav. SS Econ. Comput. Serv.*, vol. 28, no. 8, pp. 1258–1271, Oct. 2012.

[11] M. I. A. Ajlouni, W. Hadi, and J. Alwedyan, "Detecting phishing websites using associative classification," *Image (IN)*, vol. 5, no. 23, 2013.

[12] P. A. Barraclough, M. A. Hossain, M. A. Tahir, G. Sexton, and N. Aslam, "Intelligent phishing detection and protection scheme for online transactions," *Expert Syst. Appl.*, vol. 40, no. 11, pp. 4697–4706, Sep. 2013.

[13] G. Ramesh, I. Krishnamurthi, and K. S. S. Kumar, "An efficacious method for detecting phishing webpages through target domain identification," *Decis. Support Syst.*, vol. 61, pp. 12–22, May 2014.

[14] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5948–5959, 2014.

[15] D. Zhang, Z. Yan, H. Jiang, and T. Kim, "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites," *Inf. Manage.*, vol. 51, no. 7, pp. 845–853, 2014.

[16] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Syst. Appl.*, vol. 53, pp. 231–242, 2016.

[17] Y. Li, R. Xiao, J. Feng, and L. Zhao, "A semi-supervised learning approach for detection of phishing webpages," *Opt.-Int. J. Light Electron Opt.*, vol. 124, no. 23, pp. 6027–6033, 2013.

[18] G. A. Montazer and S. ArabYarmohammadi, "Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system," *Appl. Soft Comput.*, vol. 35, pp. 482–492, 2015.

[19] Y. Li, L. Yang, and J. Ding, "A minimum enclosing ball-based support vector machine approach for detection of phishing websites," *Opt.-Int. J. Light Electron Opt.*, vol. 127, no. 1, pp. 345–351, 2016.

[20] "UCI Machine Learning Repository: Phishing Websites Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites#. [Accessed: 29-Jan-2017].

[21] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," presented at the Internet Technology And Secured Transactions, 2012 International Conference for, 2012, pp. 492–497.

[22] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Phishing websites features," *Unpubl. Available Httpeprints Hud Ac Uk243306RamiPhishingWebsitesFeatures Pdf*, 2015.

[23] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Intelligent rule-based phishing websites classification," *IET Inf. Secur.*, vol. 8, no. 3, pp. 153–160, 2014.

[24] M. Hall, I. Witten, and E. Frank, "Data mining: Practical machine learning tools and techniques," *Kaufmann Burlingt.*, 2011.

[25] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[26] L. I. Kuncheva and J. J. Rodríguez, "An experimental study on rotation forest ensembles," presented at the MCS, 2007, pp. 459–468.

[27] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, 2006.

[28] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[29] N. A. Obuchowski, "Receiver Operating Characteristic Curves and Their Use in Radiology 1," *Radiology*, vol. 229, no. 1, pp. 3–8, 2003.

[30] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques.," *Invest. Radiol.*, vol. 14, no. 2, pp. 109–121, 1979.

TABLE I. PERFORMANCE OF DIFFERENT DATA MINING TECHNIQUES FOR PHISHING WEBPAGE DETECTION

	ANN	k-NN	SVM	C4.5	Random Forest	Rotation Forest
ROC Area	0.995	0.989	0.97	0.984	0.996	0.994
F – measure	0.969	0.972	0.972	0.959	0.974	0.968
Accuracy	96.91%	97.18%	97.17%	95.88%	**97.36%**	96.79%