2017 International Conference on New Trends in Computing Sciences

Phishing Websites Prediction Using Classification Techniques

Dyana Rashid Ibrahim
Dept. of Computer Science
Princess Sumaya University for Technology
Amman, Jordan
dyana.r.ibrahim@gmail.com

Ali Hussein Hadi
Dept. of Computer Science
Princess Sumaya University for Technology
Amman, Jordan
a.hadi@psut.edu.jo

***Abstract*— Phishing is an important issue that faces the cyber security. This paper exploits the capabilities of classification techniques on Phishing Website Prediction (PWP), and introduces a methodology to protect users from the attackers. The blacklist procedure isn't a strong enough way to stay safe from the cybercriminals. Therefore, phishing website indicators have to be considered for this purpose, with the existence and usage of machine learning algorithms. Five different classification techniques have been used to evaluate their efficiency on (PWP) in terms of accuracy and the Relative Absolute Error (RAE) value for each one of them, with and without the feature selection process. WEKA tool was used for the implementation of these classifiers on a public dataset from NASA repository. The motivation behind this investigation is to employ a number of Data Mining (DM) algorithms for the prediction purpose of phishing websites and compare their effectiveness in terms of accuracy and RAE. Where DM classifiers have proved their goodness in this kind of problems.**

Keywords*— *data mining; feature selection; machine learning; optimization algorithms; phishing website; prediction; supervised algorithms.

I. INTRODUCTION

Phishing website is a type of attack that aims to steal sensitive information that are important to the attacker from users. Phishing is done through sending an email to the victim to be as a bait, or through chat rooms "instant messaging" [1]. Victims get into a fake website that belongs to the attacker, believing that this is the desired one. Then the attacker steals sensitive information like credit card information and passwords.

The reason behind users being fooled may come from the good creation of similar web pages to the original ones that seem to be legitimate (having the same website architecture and design, similar domain names, etc.), the weak knowledge of technologies and the lack of awareness of possible attack techniques.

In the phishing activity report that was released on March 2016 by the Anti-Phishing Working Group (APWG), they found that the number of unique phishing websites detected was 123,555 where it was increased from 48,114 that were detected on October 2015 [2]. Therefore, many researches have been done to keep users safe from malicious websites through the process of detecting them. The structure of this paper is as follows: section 2 describes the related work, section 3 represents data mining classification techniques, section 4 represents the proposed methodology, section 5 represents the phishing websites dataset, section 6 represents the experimental results and section 7 represents the conclusion and future work.

II. RELATED WORK

In [3] study, authors made two case studies for e-banking phishing; phone phishing experiments and website phishing experiments to test the awareness of users. A number of phishing features and indicators have been concluded from these studies. A comparison between six different classification algorithms was conducted, they found out that the MCAR (Multi Class Classification based on Association Rule) algorithm had the best performance.

In [4] study, authors made a hybrid feature selection and classification approach for automating phishing email detection, feature selection is used to remove the irrelevant and redundant features that may negatively affect the classification performance. Features were extracted from the email header and body then the Hybrid Feature Selection (HSF) algorithm was used for feature selection and as the last step the Bayes Net algorithm was used for the classification purpose.

In [5] study, authors have gathered 15 phishing email features from different studies and applied the feature selection process based on the highest Information Gain (IG) of these defined features, 14 of them were selected. The Random Forest (RF) classification algorithm was applied for the email phishing detection process, this approach was tested on different dataset sizes, and they found that their work performs better on the large datasets.

In [6] study, authors have made and implemented a new framework where they have extracted a number of phishing website features from many collected phishing websites, Naïve Bayes classifier (NB) and Support Vector Machine (SVM) algorithm were used for training. Ten classifiers were used for the prediction purpose and then the ensemble method combines these prediction results to have a better performance and more accurate results, as the last step researchers perform the clustering process for the phishing categorization purpose as "Online payment", "Booking net", "Lottery ticket", etc.

In [7] study, authors used an optimization technique as a classifier and compared it's results of other classifiers, the Particle Swarm optimization algorithm (PSO) was used, researchers found that this algorithm has the best performance

978-1-5386-0527-1/17 $31.00 © 2017 IEEE
DOI 10.1109/ICTCS.2017.38

in terms of accuracy and efficiency than the other compared traditional data mining classification algorithms.

In [8] study, authors have been proposed a new approach that employs the fuzzy logic with the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm, where this integration takes the linguistic variables advantage from fuzzy logic in addition to the rule learner capability from RIPPER algorithm. They used content based and non-content based approaches characteristics for the classification purpose. After detecting phishing website, they used the WHOIS protocol to define the owner of that phishing page.

In [9] study, authors used dataset from PHISHTANK archive. They have extracted the most relevant features from the URL and HTML source code in addition to the blacklist property. Then three supervised learning algorithms (Multilayer Perceptron (MLP), NB and Decision Tree (DT) classifiers) have been used to decide whether the website is legitimate or phishing one. Decision Tree classifier had the best accuracy results.

III. Data Mining Classification Techniques

Data mining is the process of extracting useful and relevant information based on the required purpose, sometimes it's called Knowledge Discovery in Databases (KDD) [10]. Classification is a supervised machine learning technique that is applied on the already labeled data to give a predicted separation of it using different classifiers. Fig. 1. illustrates the Machine Learning (ML) classification process in general:

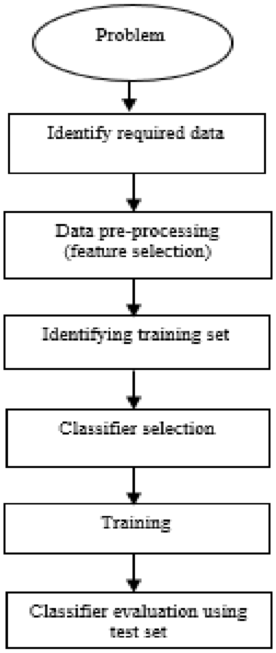

Fig. 1. The ML classification process.

There are many classification algorithms such as PRISM [11], C4.5 [12], MCAR [13], Naive Bayes (NB) [14], Artificial Neural Networks (ANN), etc.

IV. Proposed Methodology

This methodology will help on the prediction of phishing websites. The approach is illustrated in Fig. 2.

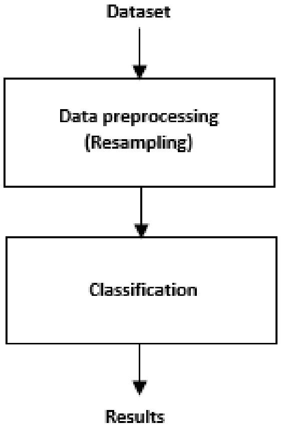

Fig. 2. General approach of Phishing Website Prediction.

The first step is the resampling process; a supervised instance resampling was done to overcome the class imbalance problem to have better accuracy, where random subsamples are produced with replacement. Then the classification process takes place to predict the class (legitimate or phishing) through a number of classifiers. Finally, the validation process in terms of accuracy and the RAE will be considered to compare the efficiency of the tested classifiers on PWP. The used classifiers are illustrated in A, B, C, D and F:

A. Prism

The Prism algorithm is based on ID3 decision tree algorithm, where it was conducted to enhance its work. Prism's strategy finds rules that are associated to attribute-values with each class. This algorithm generally works as follows; it computes the probability of each attribute value pair in each class, take the feature value with the highest probability and create subset that contains the instances with that feature value, repeat these steps on the subset and produce rules using the correlation of the features with the classes [11].

B. Multi-Layer Perceptron (MLP)

Artificial Neural Networks (ANN) works similar to brain neurons. The Single-Layer Perceptron can only classify a set of instances that can be separated linearly to their categories, but if instances cannot be separated linearly then the classification process will not be correct to all instances [12]. Therefore, the Multi-Layer Perceptron was created, which is a feedforward artificial neural network. MLP consists of input layer, hidden layer (could be one or more) and the output layer. The input layer has the training set, neurons on the hidden layers do the computational process and then we will have the results on the output layer (forward propagation from the input to the output), if an error has occurred which is the difference between the actual output and the desired one, a backpropagation process will be implemented, which means that the process will be in the reverse order (from the output to the input) in order of modifying the weights to reduce the error percentage. Then the network will be used to classify a new set of data.

C. Naive Bayes (NB)

Naïve Bayes is a statistical classifier; it depends on the concept of "class conditional independence"; where each attribute value has independent relationship with classes (each attribute has its independent probability of belonging to a class, regardless of any correlation between the features) and that's for the computation simplicity. NB uses this formula: P(X|Ci)P(Ci) to predict the class of a sample, where P is the probability, X corresponds to the set of features X = {x_1,x_2,…, x_n} and C_i represents the class. After computing the probability of each attribute, use the mentioned formula to find the probability of these attributes in each class. Finally compare these values and the class that has the highest value will be the predicted one for that sample [14].

D. KStar (K*)

K* is an instance based algorithm, where it depends on the classification process on the already existed and classified instances; it compares the new instance with the ones that are available in the database and compute the distance using K nearest neighbor algorithm to detect the most similar instance to the new one to predict its class. The probability of instance existence in a class is calculated through summing the probability of similarity between this instance with other instances that are belonging to this class. It has proved its efficiency among all instance-based learners [15].

E. Random Forest (RF)

Random Forest is a collection of decision trees; the ensemble method is used to improve the accuracy of the classification process. Where as a first step, the bagging technique is generated to choose a random number of instances that constitute a subset, then bagging technique is repeated to choose random features (bagging attributes) after specifying the number of attributes using the following equation $p = (n)^{1/2}$, where p is the number of chosen attributes and n is the number of the whole attributes. Now for each subset of instances with the randomly selected features compute the misclassification rate (confusion matrix) and choose the one with the least value, repeat this process for each subset. At the end a forest of decision trees will be generated. For the classification purpose each decision tree predict the class of the instances, then according to votes from trees the predicted class will be assigned to the instances according to the highest votes of the predicted class [16].

RF can classify large datasets with high accuracy, a fast algorithm, doesn't suffer from the problem of overfitting and it has the ability for estimating missing data.

V. PHISHING WEBSITES DATA SET

One of the problems that faces some researches that they don't use public datasets. Due to the unavailability of reliable phishing website datasets some researchers made an investigation on the most effective features that contribute on detecting phishing websites and have published it on the UCI repository, this dataset was collected on 2015, from PhishTank archive, MillerSmiles archive, and some new features were added, it has 30 attributes and 11055 instances, phishing website indicators that are in this dataset are assorted into four categories; address bar based features, abnormal based features, HTML and JavaScript based features and domain based features. Good accuracy results have been obtained from classifiers [17]. So this public dataset was used in this research.

Table 1 represents the phishing features and indicators that has been considered in the mentioned and used dataset.

TABLE I. WEBSITE PHISHING INDICATORS [17]

Criteria	Phishing Indicators
Address Bar based Features	Using the IP Address
	Long URL to Hide the Suspicious Part
	Using URL Shortening Services "TinyURL"
	URL's having "@" Symbol
	Redirecting using "//"
	Adding Prefix or Suffix Separated by (-) to the Domain
	Sub Domain and Multi Sub Domains
	HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
	Domain Registration Length
	Favicon
	Using Non-Standard Port
	The Existence of "HTTPS" Token in the Domain Part of the URL
Abnormal Based Features	Request URL
	URL of Anchor
	Links in <Meta>, <Script> and <Link> tags
	Server Form Handler (SFH)
	Submitting Information to Email

	Abnormal URL
HTML and JavaScript based Features	Website Forwarding
	Status Bar Customization
	Disabling Right Click
	Using Pop-up Window
	IFrame Redirection
Domain based Features	Age of Domain
	DNS Record
	Website Traffic
	PageRank
	Google Index
	Number of Links Pointing to Page
	Statistical-Reports Based Feature

VI. Experimental Results

WEKA tool has widely used from Machine Learning researchers, where it has proved its strength in data mining tasks, like; clustering, regression and classification, because it provides many algorithms for these tasks that can be used on our datasets. In this investigation the classification task will be examined in order to classify instances and predict their category in PWP problem; instances will be classified to legitimate or phishing website. Two cases have been tested to evaluate the classification performance; with and without feature selection. In the beginning the resampling technique was done to the instances before classification process takes its place, due to its effect in increasing the accuracy percentage. Where resampling takes out a random subsample from the dataset, through sampling with or without replacement.

Different five types of classifiers have been tested; classifiers were chosen due to their different nature of work, to compare their efficiency in terms of accuracy (correctly classified instances) and RAE in Phishing Website Prediction problem, in both cases with and without feature selection. Experiments were done using 10-fold cross validation as testing mode.

Tables 2 and 3, represents the obtained results from applying feature selection for preprocessing, before the execution of the classification techniques. Where it decreases the high dimensionality. Therefore, the performance will be increased, but as seen that it doesn't increase the accuracy of the tested classifiers.

Many metaheuristic optimization algorithms were used to do the purpose of feature selection, and it was found that they have fairly close accuracy results. The Bat Search algorithm is one of them, which had the best accuracy results with all classifiers. It selected ten indicators from the dataset, which are; Prefix_Suffix, having_Sub_Domain, SSLfinal_State, Request_URL, URL_of_Anchor, Links_in_tags, SFH, web_traffic, Google_Index, Statistical_report. Table 2 shows the accuracy results for each classifier after doing its process on these features only.

Table II. Classification accuracy results with Bat-search algorithm as feature selection

	NB	**MLP**	**K***	**Prism**	**RF**
Testing mode	10 Fold Cross Validation				
Correctly classified instances	92.7%	94.9%	94.6%	88.4%	95.2%
Incorrectly classified instances	7.3%	5.1%	5.4%	11.6%	4.8 %

Other metaheuristic algorithms have been tested, to discover their efficiency in selecting the most effective features. Therefore, check if better accuracy results are gained or not. The Particle Swarm Optimization (PSO), FireFly, Genetic, Bee and the BestFirst algorithms were tested, it was found that all of these algorithms choose the same nine features; that are the same of what have been selected from the Bat search algorithm except the Statistical report metric.

Table 3 represents the classifiers accuracy with these nine metrics. Results were relatively close to each other. RF classifier got the same accuracy results with all tested feature selection algorithms, which implies on the stability and strength of this algorithm. Prism algorithm was the most affected one, because it depends on the existence of the attributes; to make rules between them and the classes.

Table III. Classification accuracy results with PSO algorithm as feature selection

	NB	**MLP**	**K***	**Prism**	**RF**
Testing mode	10 Fold Cross Validation				
Correctly classified instances	92.6%	94.8%	94.3%	86.9%	95.2%
Incorrectly classified instances	7.4%	5.2%	5.7%	12.1%	4.8%

On the other hand, as shown in table 4, the obtained results from classifiers without applying the feature selection process has increased the accuracy but the performance was a little bit decreased. RF classifier had the best performance in terms of accuracy and the K* algorithm had a very close result, while the worst result came from performing Naïve Bayes algorithm.

TABLE IV. CLASSIFICATION ACCURACY RESULTS WITHOUT FEATURE SELECTION

	NB	MLP	K*	Prism	RF
Testing mode	10 Fold Cross Validation				
Correctly classified instances	92.9%	97.7%	98.1%	97.087%	98.4%
Incorrectly classified instances	7.01%	2.3%	1.9%	2.596%	1.6%

As we observe, all classifiers accuracy without feature selection had better results than using feature selection as a preprocess step before classifying instances. Prism classifier was the most influenced algorithm; because it has better performance when executing with the existence of all features while NB classifier didn't show a big difference.

Another measure of classifiers has been considered without applying the feature selection because they had a better accuracy results. Which is the RAE; where the MLP and Prism classifiers had the least RAE values, while NB had the highest one, as shown in Fig. 3.

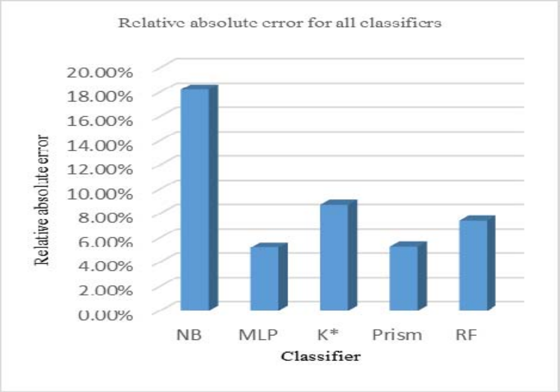

Fig. 3. Relative Absolute Error for all classifiers.

CONCLUSION AND FUTURE WORK

Classification technique showed significant performance in Phishing Website Prediction. Five algorithms have been used and compared in terms of accuracy and RAE. The feature selection preprocess was considered to observe the performance of classifiers with a minimal number of features. The obtained results showed that the classifiers are doing good without eliminating features in the tested dataset. But there is a tradeoff between the accuracy and the consumed time in the prediction process. Where if more accuracy was required, it's better to use the classification techniques without feature selection. And if the best performance was the target it's better to use the feature selection process in this dataset. Without applying the feature selection, it was found that the Random Forest classifier had the highest accuracy and the Naïve Bayes had the least one, while the Multilayer Perceptron and Prism had the least RAE value and Naïve Bayes had the highest one. In the future more classification techniques can be compared, with different measures, and more datasets can be used, in addition to the feature extraction from a number of phishing websites then we could apply many classification techniques for the prediction process.

REFERENCES

[1] P.A. Barraclough, M.A. Hossain, M.A. Tahir, G. Sexton, N. Aslam, "Intelligent phishing detection and protection scheme for online transactions", ELSEVIER, pp. 4697–4706, 2013.

[2] Anti-Phishing Working Group, Phishing Activity Trends Report, March 2016. https://docs.apwg.org/reports/apwg_trends_report_q1_2016.pdf.

[3] Maher Aburrous, M. A. Hossain, Keshav Dahal, Fadi Thabtah, "Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies," IEEE conference, 2010.

[4] Isredza Rahmi A Hamid, Jemal Abawajy and Tai-hoon Kim, "Using feature selection and classification scheme for automating phishing email detection," Studies in informatics and control, 2013.

[5] Andronicus A. Akinyelu and Aderemi O. Adewumi., "Classification of Phishing Email Using Random Forest Machine Learning Technique," Hindawi Publishing Corporation, Journal of Applied Mathematics, 2014.

[6] Weiwei Zhuang, Qingshan Jiang, Tengke Xiong, "An Intelligent Anti-phishing Strategy Model for Phishing Website Detection," International Conference on Distributed Computing Systems Workshops, 2012.

[7] R.Sumathi and Mr.R.Vidhya Prakash, "Prediction of Phishing Websites Using Optimization Techniques," International Journal of Modern Engineering Research (IJMER), 2012.

[8] Anindita Khade, Dr. Subhash K Shinde, "Detection of Phishing Websites Using Data Mining Techniques," International Journal of Engineering Research & Technology (IJERT), pp. 2278-0181, 2013.

[9] Santhana Lakshmi V, Vijaya MS., "Efficient prediction of phishing websites using supervised learning algorithms," ELSEVIER, pp. 798 – 805, 2012.

[10] Baker, R.S.J.d., "Data Mining for Education," In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), Oxford, UK: Elsevier, 2010.

[11] JADZIA CENDROWSKA, "PRISM: An algorithm for inducing modular rules," International Journal of Man-Machine Studies, pp. 349-370, 1987.

[12] S. B. Kotsiantis. "Supervised Machine Learning: A Review of Classification Techniques," pp. 249-268, 2007.

[13] F. Thabtah, P. Cowling, Y. Peng, "MCAR: multi-class classification based on association rule," pp. 2161-5330. Jan. 2005 IEEE conference.

[14] K. Ming Leung, "Naive Bayesian Classifier," 2007.

[15] John G. Cleary, Leonard E. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure," 12th International Conference on Machine Learning," pp. 108-114, 1995.

[16] Leo Breiman, "RANDOM FORESTS. Machine Learning," pp 5-32, 2001.

[17] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, " Phishing Websites Features," 2015.