ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs24822019.pdf
https://doi.org/10.30534/iiiscs/2019/24822019

Phishing Espial Using Machine Learning with Wrapper Integrant Selection and Google Classification

Bessy P Babu[1], Devika Raju[2], Karthika S[3], Keerthana Chandran[4], Jissy Liz Jose[5]
[1]Mangalam College of Engineering, India, bessypbabu@gmail.com
[2]Mangalam College of Engineering, India, devikaraju7@gmail.com
[3]Mangalam College of Engineering, India, karthikasubash16@gmail.com
[4]Mangalam College of Engineering, India, keerthanachandran2015@gmail.com
[5]Mangalam College of Engineering, India, jissyliz@gmail.com

ABSTRACT

Phishing has become the most popular practice among the criminals in internet world. The phisher creates a phishing website to obtain sensitive information from the users. We propose machine learning techniques to find the phishing websites that gives best performance than others.

This paper uses Google classification that checks the given website in Google and wrapper based feature method to extract important features that is enough to predict the phishing site correctly. Machine Learning classifiers such as C4.5 Decision Tree Algorithm, Support Vector Machine (SVM) Classification, Multilayer Perception (MLP) Artificial Neural Networks and Machine Learning with Naïve Bayes Algorithm are used as classifiers. These classifiers learn from the given dataset and match with the features of future data and predict the result as whether the website is phishing or not. Google classification and wrapper feature selection helps to improve the efficiency and processing speed of the system.

Key words: Phishing; Machine learning; Google classification; classifiers

1. INTRODUCTION

In this digital world, security is major issue faced by all the internet users and hence the security features of each system must be upgraded to a new level. Phishing is a cyber-attack which causes the theft of sensitive information from users. Phishing sites are actually mimicking the original site which looks exactly the original site. Usually the links of the phishing sites are send to email by the phishers. Users are not able to recognize whether the site is phishing or not. Phishing attack is performed by entering the details in the login form of the website. Eventually the attacker can misuse the details provided by the users. The phishing websites has been targeting several sectors like online business, bank, and online payment system. Spear phishing, whale phishing are the types of phishing attacks.

Machine learning is one of the most advanced techniques which allows computer to learn from the given data. Training classifier with certain features and these features are extracted from the dataset. Classifiers will train using various training dataset so that it can predict the output for the future input data. While training a classifier, we have to be careful on providing these features for training since prediction will depends on the given training data. In this paper, features of phishing websites will allow to learn the classifiers so that it can recognize that the user trying to use is phishing site or not.

There is so many attributes to train our machine to detect phishing. But there is a difference in accuracy on determining the phishing site that depends on the dataset used to train and the type of classifier we choose.

We uses java language to implement as it provides debugging ease, package services, graphical representation of data and better user interaction.

2. BACK GROUND

Phishing becomes a major threat; machine learning can solve these issues in a better way since Security is the fundamental issue in the internet.

On discussing phishing page detection, features extracted from the dataset and apply to a classification model. In this [1] technique first convert phish pages into 12 features. The training set including normal and phishing pages which are then given to a support vector machine to perform

training. The results show that the phishing detector can achieve the high accuracy rate with low false positive and low false negative rates.

In case of designing a hybrid model first do the individual classification according to high accuracy [2] and performance and select best three models. Then combine the weak models together and we can find that they can perform better than they did before individually. In this achieved high accuracy of combining the models and thus producing a hybrid model.

One of the most effective techniques that can be used in predicting phishing attacks is based [3] on data mining, the "induction of classification rules" since anti-phishing solutions aims to predict the website class appropriately and that exactly match with the data mining classification technique. The important features that distinguish phishing websites from legitimate ones are used to identify this phish sites.

One way to ensure the reliability of the results and to enhance performance of the phishing detection is to identify a set of feature. [4] A new feature selection method that combines the scores of multiple known methods is used and that minimize errors in feature selection results. This method has been applied to the problem of website phishing classification to show its advantages and disadvantages in identifying relevant features. A result of the security dataset tells that the preprocessing method employed was able to derive new features datasets. To generate high competitive classifiers with reference to detection rate when compared to results obtained from other features selection methods.

This research employs approach that uses fuzzy logic with classifiers like SVM, NMC and Gaussian. Fuzzy based detection [6] system provides effective aid in detecting phishing websites. It successfully resulted in low false positive and high true positive for classifying phishing website.

3. FAMILARISATION OF CLASSIFIERS

1) C4.5 Decision Tree Algorithm

C4.5 decision tree algorithm also called statistical classifier. It is like ID3 algorithm. Here the internal nodes are representing the attributes and its branches are the values for it. Leaf nodes are the decision made by the tree. It will generate decision trees from the given dataset. This will generate the decision tree as result. It uses the concept of information gain for classification of attributes. The highest information gain will normalized to take a decision [8].

2) Support Vector Machine (SVM) Classification

It is a supervised learning model for linear and non- linear data. This will find the best optimal hyper plane or hyper plane with maximum margin for the classification. Depends on the number of features, dimension of hyper plane will varies [6].

3) Multilayer Perception (MLP) – Artificial Neural Networks

It is a feed forward network made up of more than one perceptron or neurons mostly applied to supervised learning. The training occurs through the back propagation algorithm. There is a set of input- output pairs and it will compare the obtained output and expected output. Here weight, bias and parameters are adjusted to reduce error [9].

4) Machine Learning with Naïve Bayes Algorithm

It can be efficiently performs with supervised learning and works well in complex real world applications. It will work well without the Bayesian probability or any Bayesian methods if we have maximum likelihood [7].

4. PROPOSED SYSTEM

The following figure describes the working of phishing detection using wrapper feature selection and Google classification

The steps as follows

1) Google classification

It is the first step, we will search the given link in Google whether that link is in Google or not. If it available, then we go for the next step. Google ranking systems sort through billions of webpages and gives relevant and useful results within fraction of a second. These ranking systems designed with series of algorithms to analyze what the people really wants and the results they got. To rank useful webpages, Google use PageRank algorithm. For estimation of importance of websites, PageRank counts the number and quality of link to a page. A trusted website receives more links from other websites.

2) Collection of features of phishing and non -phishing websites

For the analysis of machine learning classifiers, in this paper UCI machine learning repository is uses dataset for the phishing sites. Students, educators, and researchers use these data sets for various purposes.

Collection of features of legitimate and illegitimate websites is in Attribute Relation File Format (ARFF) so that it

will be suitable for data mining. Evaluates output based on the features available from the UCI machine learning repository.

3) Wrapper feature selection

This is a preprocessing step where selecting the best features from the dataset then trains the classifier. So that it can easily separate phishing and non – phishing websites. Wrapper feature enables machine learning algorithm to train easier as it only extract the best features that classifies data correctly. Also it reduces the complexity of a model. If we choose the right subset, it improves the accuracy and will reduce the over fitting.

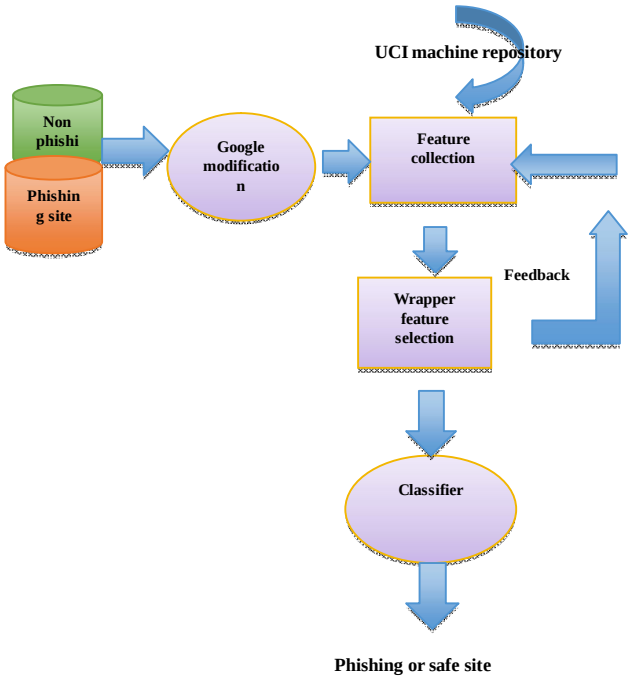

Figure 1: Architecture

4) Training the classifiers

Selected features are used to train the classifiers. Supervised learning is used here. From the given data it will learn that how a phishing site will look like. Each classifier analyses the input data and check with the phishing site features. If it matches, it will predict the output as phishing site

4) Performance evaluation of classifiers

In this phase, evaluates the classifier's accuracy and other performance measurements. If there is any undesired output from the actual it will train again until it matches with the desired output.

5. RESULTS

The dataset for the classification taken from the UCI machine repository. Phishing detection uses Machine Learning classifiers such as C4.5 Decision Tree Algorithm, Support Vector Machine (SVM) Classification, Multilayer Perception (MLP) Artificial Neural Networks, and Machine Learning with Naïve Bayes Algorithm.

To evaluate machine learning algorithm we use true positive rate (TPR), false positive rate (FPR), precision, recall, F–measure, ROC (Receiver Operating Characteristics) area, Matthews Correlation Coefficient (MCC).

As experimental result, it is found that accuracy of test is higher for both naïve bayes and SVM and but in multilayer perceptron, the F-measure has large change in the accuracy of detecting non phishing sites from the accuracy of detecting phishing sites. MCC is higher for SVM among other classifiers. True positive rate is higher for naïve bayes and also for SVM and for J48 decision tree (Java implementation of c4.5 algorithm).

	Predicted	**Predicted**
Actual	True positive	False negative
Actual	False positive	True negative

Table 1: Confusion matrix

Measure name	**Formula**
True positive rate	TPR=TP/TP+FN
False positive rate	FPR=FP / (FP + TN)
Precision	TP/TP+FP
Recall	TP/ PT+FN
F- measure	(1/recall+1/precision)/2

Table 2: Measurements for the evaluation of performance

5.1 Performance Measurements

True positive rate: Ratio of number of positives predicted as correctly to the total number of actual positives.

False positive rate: Ratio of number of positives wrongly predicted to the total number of actual negatives.

Precision: Number of predicted as positives to the number of positives predicted by the classifier.

F-measure: It is a harmonic mean between precision and recall. Its maximum value is 1 and minimum is 0. It measures how much accurate the classifier is.

Matthews correlation coefficient (MCC): Used to evaluate the quality of two class problems. It gives the values between -1 and +1 .

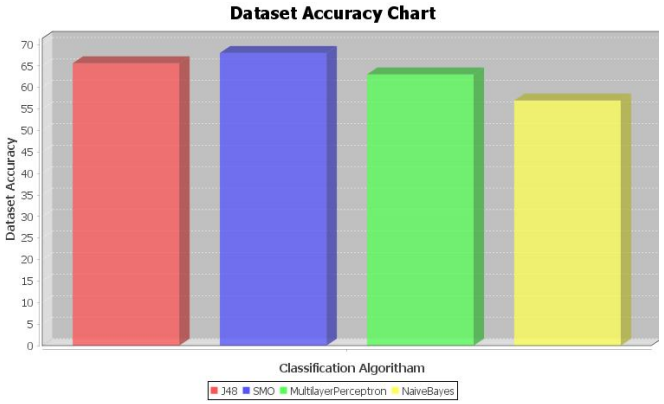

Figure 2: Comparison of accuracy of machine learning classifiers

From the accuracy chart we can found that based on our dataset, SMO support machine vector has the best dataset accuracy. It classifies the future data with accuracy of 68.01%. The worst classifier is naïve Bayesian classifier which has accuracy of 56.97%.

From observation it is clear that support vector machine give more accuracy than other classifiers. Checking the website‘s URL in Google make task of detecting phishing website easier. After the selection of wrapper features, it also reduces the processing time and increases the efficiency.

6. CONCLUSION

In this paper, we used wrapper –based features selection method and Google classification method to improve phishing detection efficiency. In order to detect the phishing websites, MLP-ANN SVM, C4.5 were applied with significant features selected by the wrapper based feature selection. The wrapper based feature selection and Google classification methods together used to improve the efficiency of detection of phishing websites by the classifier.

It is found that SVM is the best classification algorithm according to the training dataset. It shows best performance when evaluate with machine learning performance measures. C4.5 algorithm and multilayer perceptron are also good in performance. Naïve bayes is gives accuracy more than 50% but not more than among the other algorithms.

REFERENCES

1. M. He, S.J. Horng, P. Fan, M.K. Khan, R.S. Run, J.L. Lai, R.J. Chen, Sutanto, “An Efficient Phishing Webpage Detector,” Expert Systems with Applications, vol. 38(10), pp. 12018- 12027, 2011. https://doi.org/10.1016/j.eswa.2011.01.046

2. M. A. U. H. Tahir, S. Asghar, A. Zafar, S. Gillani, “A Hybrid Model to Detect Phishing- Sites Using Supervised Learning Algorithms,” International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1126-1133, IEEE, 2016.

3. R. M. Mohammad, F. Thabtah, L. McCluskey, “Intelligent Rule-based Phishing Websites Classification, “ IET Information Security, vol. 8(3), pp. 153-160, 2014. https://doi.org/10.1049/iet-ifs.2013.0202

4. Qabajeh, F. Thabtah, “An Experimental Study for Assessing Email Classification Attributes Using Feature Selection Methods,” 3rd International Conference on Advanced Computer Science Applications and Technologies (ACSAT), pp. 125-132, IEEE, 2014 https://doi.org/10.1109/ACSAT.2014.29

5. M. He, S.J. Horng, P. Fan, M.K. Khan, R.S. Run, J.L. Lai, R.J. Chen, Sutanto, “An Efficient Phishing Webpage Detector,” Expert Systems with Applications, vol. 38(10), pp. 12018- 12027, 2011. https://doi.org/10.1016/j.eswa.2011.01.046

6. Durgesh K. Srivastava, Lekha Bhambhu “Data classification using support vector machine” Journal of Theoretical and Applied Information Technology.

7. Harshad M. Kubade “The overview of bayes classification methods” vol.2 issue 4

8. Harvinder Chauhan, Anu Chauhan “Implementation of decision tree algorithm c4.5” International Journal of Scientific and Research Publications, vol. 3, Issue 10, 2250-3153 October 2013

9. Vo Hoang Trong” Decision Tree versus Multi-Layer Perceptron in Classification Problems”

10. A.Naga Venkata Sunil, Anjali Sardana “A page rank based detection techniques for phishing sites “