Detection of Phishing Websites Based On Feature Classification Protocol and Extreme Learning Machine

Miss Sneha Shivram Mande [1], Prof. K. N. Shedge [2]
Department of Computer Engineering
[1,2] Sir Visvesvaraya Institute of Technology - Nasik

Abstract- *Phishing website tries to gain the victim's confidential data by diverting them to surf a fake website page that resembles an honest to goodness one a type of criminal acts through the internet. It is one of the most common but the most hazardous attacks among all the cyber crimes. Phishing site detection is unpredictable and aims to steal the information used by individuals and organization to conduct transaction. We present a judgment model for detecting fake web sites which are determined by feature classification protocols. Web pages differ with the feature set and thus, we use it as our prime weapon to prevent the phishing attacks. The incitement of the proposed work is to perform Extreme Learning Machine (ELM) based classification for various features of the website. We thus come up with the model which uses machine learning techniques for detecting phishing web sites. ELM classification algorithm has a high rate of accuracy of detecting phishing websites as compared to other machine learning algorithms.*

Keywords- Phishing; Extreme Learning Machine; Feature Classification

I. INTRODUCTION

Technology is growing day-by-day and with this rapidly growing technology internet has become a crucial part of human's daily activities. Use of internet has grown due to the growing technology and major use of digital systems and thus, data security has gained importance. The prime goal of maintaining security in information technologies is to ensure that the necessary precautions are taken against dangers and threads which may be faced by users while using these technologies. Phishing strives to obtain sensitive information of the user and thus tries to extract the user details like passwords, credit card information etc. It pertains to be trustworthy body in an electronic communication for gaining the user information. Phishing is thus carried out by the means of internet which directs the users to enter personal information at a fake website, which lookalike the legitimate site. Phishing utilizes spoof messages which in act to be a valid one and pertains to be originated from honest to goodness sources like money related foundations, ecommerce destinations etc and divert the clients to visit fake sites through the links given in the phishing email. Phishing diverts the end users to visit fake websites and enter personal and sensitive information. The prime objective of online security is to protect the people from the fake websites and the phishing attacks. There are several methods that can be used to develop a fake web page and for this reason people using internet should be aid for securing their information and being cheated.

Many research studies have been conducted to predict the phishing websites by various means of artificial intelligence. The paper presents the study of detecting the phishing websites by extracting the features of the web sites.

II. LITERATURE SURVEY

With the consistent study in the fields of communication and information, new information security threads have come into picture. It is thus essential to prevent the individual or the institute from the damage by securing the sensitive data on the internet. There are various studies conducted for detecting the phishing attacks and it is observed that with the help of ML technique we can obtain high level of accuracy in the detection of phishing websites.

Vijaya and Santhana Lakshmi[3] together came forward and proposed a model which used supervised Machine-learning technique for predicting the wok. They used Naïve Bayes classification algorithm and Decision tree. It was further observed that the decision tree prediction was more accurate than the other algorithms used in those days.

Li et al.[5] experimented and came up with the method which used ball support vector machine (BVM) which helped in detecting phishing website. Study was carried to extend the feature vectors and they thus proposed a method that extracted the feature set of the examined web pages. The proposed method used SVM technique as a classifier which has two phases, the first one is the training phase and the second is the testing phase. While the training phase, the method extracts feature set and during the testing phase it

predicts whether the website is either phishing or legitimate website. It aimed to achieve not only high speed but also greater accuracy while detecting the phishing sites

Chen et al[6] along with his fellow members experimented and evaluated the phishing identification by the means of the risk levels of the targeted companies and the market value losses. All together, the phishing alerts analyzed was around 1030 alerts. The experimental method predicted 89% accuracy in terms of severity of the attack while it used text supervised classification and phrase extraction methods.

III. PROPOSED METHODOLOGY

The Proposed system is based on real-time phishing detection and a machine learning process. Mostly the phishing URLs have couple of connections between the part of the URL which means inter-relatedness and by using these features the phishing URLs are extracted. Then the extracted features are further used for a machine-learning classification and thus to detect phishing websites on real time. We defined some protocols and gave equations of web features. We use equations in order to explain phishing attacks characterization.

Finally, we discuss the application layout, architecture, and the communication that takes place within different modules and identify the website feature classification. Also, we intend to present a general model of an application which we will refer throughout the thesis. This helps us to study our problem from general perspective.

A. Architecture

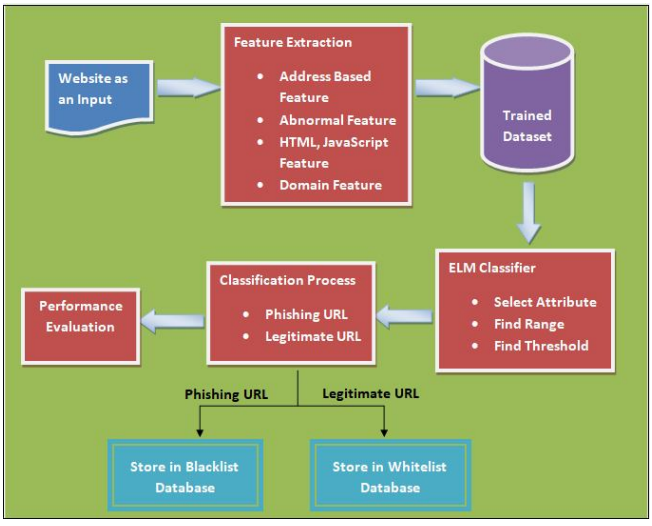

Fig. 1. System architecture and flow

The proposed methodology imports dataset of legitimate and phishing URLs from the database and then the data is preprocessed. Detecting phishing website is performed based on following four categories of URL features: address based, domain based, abnormal based and HTML, JavaScript features. The URL features are extracted with processed data and value of the URL attribute is generated.

The inspection of the URL is performed by machine learning technique which calculates range value and the threshold value for various attributes of the URL. The URL is then classified into legitimate and phishing URL. The attributes are analyzed using feature extraction of phishing websites and it is thus used to identify the range value and the threshold value. The value of the phishing attributes ranges from {-1, 0, 1} these values are defined as low, medium and high according to phishing website feature. The classification of legitimate and phishing website is based on the of the attributes extracted using phishing categories and a machine learning approach.

URL Feature Analysis

The phishing attribute features are extracted from the URL. Using this feature attribute we can find whether the website is phishing or legitimate.

Following are the protocols created for examining the data set for the extracted feature attribute:

Feature 1 – By the IP address:

If IP address exists in the domain →Phishing URL, else → Legitimate URL

Feature 2 – By the URL length:

If the length of URL < 54 → Legitimate URL, else if the length of URL ≥ 54 but ≤ 75 → Suspicious URL, else → Phishing URL

Feature 3 - By tiny URL:

If tiny URL is used → Phishing URL, else → Legitimate URL

Feature 4 – URL consisting of "@" symbol:

If the URL has @ symbol → Phishing URL, else → Legitimate URL

Feature 5 – Re-directing with "//" symbol:

If in the URL, the "//" symbol is located > 7th position → Phishing URL, else → Legitimate URL

Feature 6 – By adding "-" symbol:

If domain name has "-" symbol → Phishing, else → Legitimate

Feature 7 – By Sub-domain and the multi sub-domain:

If the total number of dots in a domain section = 1 → Legitimate URL, else if the total number of dots in domain section = 2 → Suspicious, else → Phishing URL

Feature 8 – By considering HTTPS:

If uses HTTPS, issuer is trusted certificate providers and age of certificate ≥ 1 year → Legitimate URL, else if uses HTTPS, issuer is un-trusted certificate providers → Suspicious, else → Phishing URL

Feature 9 – By considering domain registration length:

If a domain expires within 1 year → Phishing URL, else→ Legitimate URL

Feature 10 – By considering favicon:

If favicon loaded from internal domain → Legitimate URL, else if the favicon loaded from external domain → Phishing URL

Feature 11 – By port status: Following are the standard port

numbers and its preferred statuses: port 21 as closed, port 22 as closed, port 80 as open, port 443 as open, port 445 as close and port 3389 as close.
Rule: If port number belongs to preferred status → Legitimate URL, else→ Phishing URL

Feature 12 - By considering HTTPS token:

If HTTPS token is not a part of domain of URL → Legitimate URL, else → Phishing URL

Feature 13 – By the requesting URL:

If percentage of request URL is less than 22% → Legitimate URL, else if percentage of request of URL is greater than or equal to 22% but less than 61% → Suspicious, else → Phishing URL

Feature 14 – By URL of the anchor:

If the percentage of URL of anchor is less than 31% → Legitimate URL, else if the percentage of URL of anchor is greater than or equal to 31% but less than or equal to

Feature 15 – By the links in the tags:

If percentage of links in <meta>, <script>and<link>tags is less than 17% → Legitimate URL, else if % of links in <meta>, <script> and <link> tags is greater than or equal to 17% but less than or equal to 81% → Suspicious URL, else → Phishing URL

Feature 16 – By considering Server Form Handler (SFH):

If SFH has an empty string or "about: blank" → Phishing URL, else if SFH belongs to a different domain → Suspicious URL, else → Legitimate URL

Feature 17 – By submitting details to e-mail:

If "mailto:" or "mail ()" functions used → Phishing URL, else → Legitimate URL

Feature 18 – By an abnormal URL

If the host name is present in URL → Legitimate URL, else → Phishing URL

Feature 19 – By website forwarding method:

If number of redirect page is less than or equal to 1 → Legitimate URL, else if number of redirect page greater than or equal to 2 but strictly less than 4 → Suspicious, else → Phishing URL

Feature 20 – By customizing status bar

If on Mouse Over the status bar remains unchanged → Legitimate URL, else if the status bar changes → Phishing URL

Feature 21 – By disabling right click event:

If right click enabled → Legitimate URL, else → Phishing

Feature 22 – By pop-up window:

If popup window does not contains text field → Legitimate

else → Phishing URL

Feature 23 – By Iframe redirection:

If iframe is not used → Legitimate URL, else → Phishing URL

Feature 24 – By considering domain age:

If domain age is less than 6 months → Phishing URL, else
→ Legitimate URL

Feature 25 – By considering DNS:

If DNS record exist for domain → Legitimate URL, else → Phishing URL

Feature 26 – By website traffic:

If Website Rank is less than 100,000 → Legitimate URL, else if the Website Rank is greater than 100,000 → Suspicious URL, else → Phishing URL

Feature 27 – By PageRank:

If PageRank > 0.2 → Legitimate URL, else → Phishing URL

Feature 28 – By Google Index:

If the web page is indexed by Google → Legitimate URL, else→ Phishing URL

Feature 29 – By considering links pointing the page:

If total number of links pointing a webpage is equal to zero → Legitimate URL, else if total number of links pointing, To webpage is greater than zero but less than or equal to two → Suspicious URL, else → Legitimate URL

Feature 30 – Feature based on statistical reports:

If the host is not enlisted in top 10 phishing IPs or domains → Legitimate URL, else → Phishing URL

B. Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) is proposed as a single hidden layer feed-forward artificial neural network (ANN) model which ensure a high-performing learning and parameters such as threshold value, weight and activation the function must have appropriate values for the data system which is to be modeled. In ELM learning, the parameters are gradient-based, where the input weights are randomly selected while the output weights are analytically calculated. For the sake of activating the cells in the hidden layer of ELM, a linear function as well as non-linear (sinus, sigmoid, Gaussian), and the non-derivable or discrete activation functions can be used.

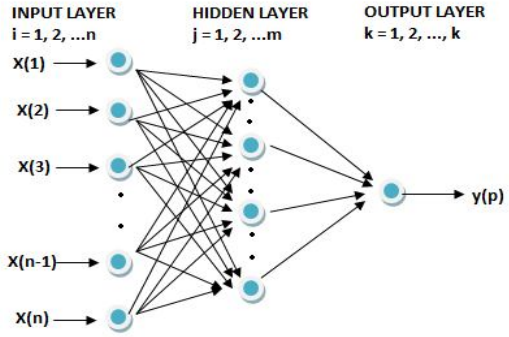

Fig. 2. ELM network model

Here, n: training samples, m: number of classes, i = 1, 2,…, n, j = 1, 2, …m, k = 1, 2, …, k, xi: input vector and y(p): desired output vector.
There are three layers, which are input layer, the hidden layer and the output layer.

$$y(p) = \sum_{j=1}^{m} \left[\beta j \; a\left(\sum_{i=1}^{n} [wi,j.xi+bj] \right) \right] \quad (1)$$

In above equation 1, wi,j is an input layer to hidden layer weights and βj is an output layer to hidden layer weights, bj is the threshold value of neurons in the hidden layer and a(.) is the activation function. In the input layer, weights (w) and bias (bj) values are randomly assigned in the equation. The activation function (a(.)), input layer neuron count (n) and hidden layer neuron count (m) are assigned in the beginning

C. Algorithms

Step1: Enter a URL of a website.

Step 2: Examine all the attributes of the website or the web page according to its features.

Step 3: Fetch all the samples features to the dataset.

Step 4: Randomly select 10% of the testing samples while 90% training samples of the dataset.

Step 5: Apply ELM classification algorithm on the dataset

Step 5.1: Arbitrarily generate hidden node parameters.
Step 5.2: Calculate output matrix for the hidden layer.
Step 5.3: Calculate weight of the output matrix.

Step 6: Prediction for website whether phishing or l.

D. K-Fold Validation Test

K - Fold validation method is used to evaluate the machine learning models where the data sample is limited and unseen. As a result of the operation, there are two methods to measure the performance of the model and the algorithm. We can measure the accuracy of the system and thus evaluate its success by using the validation test. Firstly we divide the dataset into training set and the test data set, and secondly we apply k-fold cross validation test. K-fold cross validation method is a regression test method which is applied on all the sub-sets of the given dataset. The overall success of the system is measured by considering the average calculation.

IV RESULT AND DISCUSSIONS

In the experimental study, database is creates for phishing websites and are classified by simply determining the input and the output parameters for the ELM classifier. The result obtained by ELM classifier has greater accuracy achievement as compared to the other classifiers i.e. Support Vector Machine (SVM) and Naïve Bayes (NB) methods. The study is thus considered to be an applicable design with high performing classification against the hazardous phishing activity of the websites. Also, if we compare the literature study the proposed study is observed to be high-performing this has greater accuracy of 92.18% which is also the highest accuracy in the publication.

The data set is collected from Google search operators and PhishTank archive. The major challenge faced while the study was the lack of reliable training datasets, and this problem is faced by all the researchers who have worked and study in this area.

Training dataset and phishing website features were used from the source [14] for the study. There are values and attributes for the input and output dataset. Input dataset has 30 attributes while the class in the output dataset takes value as 1 or -1. The output dataset obtained may take two different values.

TABLE I. Accuracy of Classification

Classification Method	Train Accuracy	Test Accuracy
Extreme Machine Learning (ELM)	100%	95.34%
Support Vector Machine (SVM)	100%	93.80%
Naïve Bayes (NB)	100%	92.98%

The aim of the application is to determine the types of attacks that cyber threats called as phishing attacks. Extreme Learning Machine classification algorithm is used for this intended purpose.

V. CONCLUSION

The purpose of the application is to determine the phishing attack, thus the paper defines the various features of phishing attacks. We have proposed a classification model so as to classify the phishing attacks and it consists of feature extraction from the given website. We have defined set of protocols of phishing feature extraction which thus helps us in extracting the features. Extreme Learning Machine classification algorithm is used in order to classify the features. The result of our study gets to classify the websites and brings the highest average accuracy score of 95.34% when compared to SVM algorithm and NB classification algorithm.

ACKNOWLEDGEMENT

I am colossally gratifying to Dr. K. T. V. Reddy, Principal, Sir Visvesvaraya Institute of Technology (SVIT), Nashik for inspiring me towards this and for implausible backing and leadership too. As well we prolong obligations towards Prof. Shedge K. N. (Asst.Professor), HOD M. Tech (CSE) of Computer Engineering Department, Prof. Thosar D. S., Assistant Professor and M.E Coordinator and Staff Members for their appreciated Assistance and Provision.

REFERENCES

[1] Mustafa KAYTAN and Davut HANBAY, “Effective Classification of Phishing Web Pages Based On New Rules By Using Extreme Learning Machines” , Anatolian Journal of Computer Sciences, Vol:2 No: 1, pp: 15-36, 2017

[2] Yasin Sonmez, Turker Tuncer, based Huseyin Gokal, Engin Avci, “Phishing Web Sites Features Classification Based On Extreme Learning Machine ” IEEE 2018 6th International Symposium on Digital Forensic and Security (ISDFS), DOI: 10.1109/ISDFS.2018.8355342

[3] V. Santhana Lakshmi and M. Vijaya, "Efficient prediction of phishing websites using supervised learning algorithms", Procedia Engineering, 30, pp.798-805, 2012.

[4] 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.

[5] Y. Li, L. Yang "A minimum enclosing ball-based support vector machine approach for detection of phishing websites", Optik, 127(1), pp.345-351, 2016.

[6] X. Chen, I. Bose, A. C. M. Leung and C. Guo, "Assessing the severity of phishing attacks: A hybrid data mining approach", Decision Support Systems, 50(4), pp.662-672, 2011.

[7] W. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," Appl. Soft Comput. J., vol. 48, pp. 729–734, 2016

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

[9] 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.

[10] Phishing Activity Trends Report, Anti Phishing Working
[11] Group (APWG), 1st-3rd Quarters 2015.

[12] Internet: https://en.wikipedia.org/wiki/PageRank.

[13] Internet: http://who.is/

[14] Internet: http://www.phishtank.com