Detection of Phishing Web Sites Based On Extreme Machine Learning

Miss Sneha Mande[1], Prof.D.S.Thosar[2],

[1] Student, Computer Engineering Department, Sir Visvesvaraya Institute of Technology, Chincholi, Nashik, Maharashtra, India
[2]Assistant Professor, Computer Engineering Department, Sir Visvesvaraya Institute of Technology, Chincholi, Nashik, Maharashtra, India

ABSTRACT

Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites.

Keyword: - *Extreme Learning Machine, Features Classification, Information Security, Phishing.*

1. Introduction:

Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that the end goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers. In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavors in messages and identifying phishing substance on sites, phishers think of new and half breed strategies to go around the accessible programming and systems. Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the features of a dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to be originating from honest to goodness sources like money related foundations, e-commerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email.

1.1 Phishing Detection

In Phishing E-mail Detection Based on Structural Properties [1], the proposed approach explains to find phishing through appropriate identification and usage of structural properties of email. The experiment is done by SVM and classification technique to classify phishing e-mails. The technique is used to identify phishing e-mails, which is low in efficiency and scalability. This is purely based on structural properties of e-mail and it has to extend more structural or content properties to reduce error results. Identifying phishing target based on semantic link network [2], the paper proposes a novel approach to discover phishing website by calculating association relation among webpages that include malicious webpages and its associated webpages to measure the combination of text relation, link relation and search relation. The semantic link network proposes a strategy based on situations to identify the suspicious webpage as phishing. The disadvantage in this approach is more kind of association has to be done, similarities between visual, layout and domain has to be related. This method is considered as a time consuming approach and also various sub-relations in the combined association relations are studied. Fuzzy Neural Network for Phishing Emails Detection deals with phishing email. It distinguishes phishing email and ham email in online mode. It is adopted on rank fetching, feature fetching and grouping similar features of email. The technique is based on binary value 0 or 1 to produce the result for all features used in this method, where 1 denotes a phishing feature and 0 for non-phishing. This technique does not have much dynamic system and thus it is inadequate in performance to produce accurate results. Intelligent Phishing Website Detection and Prevention System proposed a system using link guard algorithm and works for hyperlinks. The algorithm performs tests like comparison of the DNS of actual and visual links, pattern matching, checks dotted decimal of IP address and checks encoded links. The disadvantage of this system are, it produce the incorrect positive results if any genuine site has IP address instead of domain name, and it considers few phishing site as normal if the user does not happen to visit the original site. This results in false negative conclusions. In Said Afroz, Rachel Greenstadt - Phishzoo Approach [5], the algorithm detects current phishing sites by matching their content with genuine site. This will match structure, contents and the images of website with trusted one in order to avoid phishing. Drawback of this algorithm is, it requires matching image site and it is less robust for detecting phishing attacks.

1.2 Proposed System

The Proposed algorithm is based on automated real-time phishing detection and a machine learning process. The phishing URLs mostly have connections between the part of the URL which means an inter-relation and by using it the features of phishing URLs are extracted. Then the extracted features helps real time detection of phishing websites using machine-learning classification.

2. Literature Survey

In recent years, the Web are availability of numerous services such as online banking, social networking, entertainment, education and software downloading. Accordingly, a huge volume of knowledge is downloaded and uploaded constantly to the Web. This gives opportunities for criminals to hack important personal or financial information, such as user credentials, account numbers and national insurance numbers. This is a Web phishing attack, which is the major problems in Web security [1], [2]. In a Web phishing attack, phishing websites are created, which are similar to the legitimate websites to deceive Web users so as to obtain their crucial financial and personal information. The phishing attack is performed through clicking a link received via emails. Victims receive an email containing a link to validate or update their information. If this link is accessed by the target victims, the browser will redirect them to a phishing website that appears similar to the original website. The hacker can then steal the important information of the web users, since they are asked to input the sensitive information on the phishing website. Thus, the attackers can carry out theft after phishing occurs [3]-[5]. Due to the inevitability of phishing websites targeting online banks, businesses, web users, and government, it is necessary to prevent Web phishing attacks in the early stages. While, detection of a phishing website is a challenging task, due to the many innovative methods used by phishing hackers to deceive web users [6]-[8]. The success of phishing website detection techniques mainly depends on identifying phishing websites correctly and within an acceptable timescale [2], [4]. Most of the conventional techniques based on fixed black and white listing databases have been suggested to detect phishing websites. The techniques are less efficient enough, since a new website can be launched within few seconds. Thus, most of these techniques fail to make an accurate decision dynamically on whether the new website is phishing or not. Most new phishing websites may be classified as legitimate websites [1], [2], [6]-[8]. As alternative solutions to the conventional phishing website detection techniques, some intelligent phishing detection methods have been developed and suggested in order to effectively catching phishing websites. In recent years, the supervised machine learning techniques have become common, which are smart and adaptive to the Web environment compared to the conventional phishing website detection methods.

He et al. [6] proposed a phishing detection scheme using a support vector machine based on 12 features. Barraclough et al. [7] used a Neuro-Fuzzy scheme with five inputs (User-behavior profile, Legitimate site rules, PhishTank, User-specific sites, Pop-Ups from emails) to detect phishing websites with great accuracy in real-time. Mohammad et al. [9] proposed rule-based data mining classification techniques with 17 different features to distinguish phishing from websites. Mohammad et al. [4] proposed a model for predicting phishing attacks based on self-structuring neural networks. Abdelhamid et al. [1] developed an technique called Multi-Label Classifier based Associative Classification (MCAC) to detect phishing websites. Neural network (NN), support vector machine, (SVM), naïve Bayes (NB), decision tree and other classification techniques have been employed in detection of phishing websites [5], [8], [10]-[13].

In [2] the study few approaches including Support-Vector-Machine, decision-trees, rule-based techniques and Bayephphishingechniques in identifying phishing emails. A random forest algorithm is executed in PILFER (Phishing Identification by Learning on Features of Email Received) which succeeded in identifying 96% of the phishing messages with a false-positive rate of 0.1%. Email's elements are utilized as a part of the experimental results those are IP address URLs, Age of Domain, Non-coordinating URLs, "Here" Link, HTML messages, Number of Links, Domains, Number of Dots, Containing Java script, Spam-channel Output. With respect to phishing, A. Bergholz et el. [3] exhibited a technique for enhancing learning models for identifying phishing messages by feature selection. A subset of components is chosen by a wrapper method in which the purported best-first pursuit calculation efficiently adds and subtracts features to a present subset utilizing the classifier as a feature of the evaluation function.

Pradeep and Ravendra [4] suggested a model that can detect a website is phishing or not. It uses different machine learning based classification algorithm named Naïve Bayes, J48, SVM, Random Forrest, Tree Bag and IBK lazy classifier with classification accuracy respectively. Their ratio is 70-30. Where 70% comes to training and rest for testing. Our aim is to proceed their work to gain more classification accuracy using those algorithms and also we introduced a new classification algorithm named Neural Net in this experiment.

2.1 Related Work

- **Blacklist and whitelist approach**

  Blacklist and whitelist based approach: It is based on the blacklist or whitelist to identify if the currently visited website is either a phishing or legitimate website respectively. The main disadvantage of the blacklist and whitelist based approach is that it cannot differentiate the newly created phishing websites from legitimate websites.
- **Intelligent heuristics approach**: In this approach, some features of websites are gathered and are evaluated to select the most influential website features, which thus play an important role in detecting the phishing websites. The significant features of most of the websites can be utilized as training dataset. Then, the machine learning techniques are trained based and prepared training dataset in order to effectively classify the websites as either phishing or legitimate. After verifying the performance, the trained classifiers have the ability to correctly detect the new phishing websites in the real time, which may have been unseen in the training phase. Therefore, the intelligent heuristics-based approaches are able to effectively detect newly created phishing website.

- **Machine Learning**

  Machine learning focuses on developing the computational algorithms and induces patterns and rules from externally supplied instances and priori data in order to produce general models, which can make predictions about future. The machine learning is called supervised learning if known labels are given with instances in the training phase, whereas instances are not labeled in unsupervised machine learning. Many supervised algorithms have been successfully employed in different applications [19]-[20]. However, we focuses on some popular machine learning techniques such as back-propagation neural network (BPNN), support vector machine (SVM), naïve Bayes classifier (NB), decision tree (C4.5), random forest (RF), and k-Nearest neighbor (kNN).

3. Proposed System

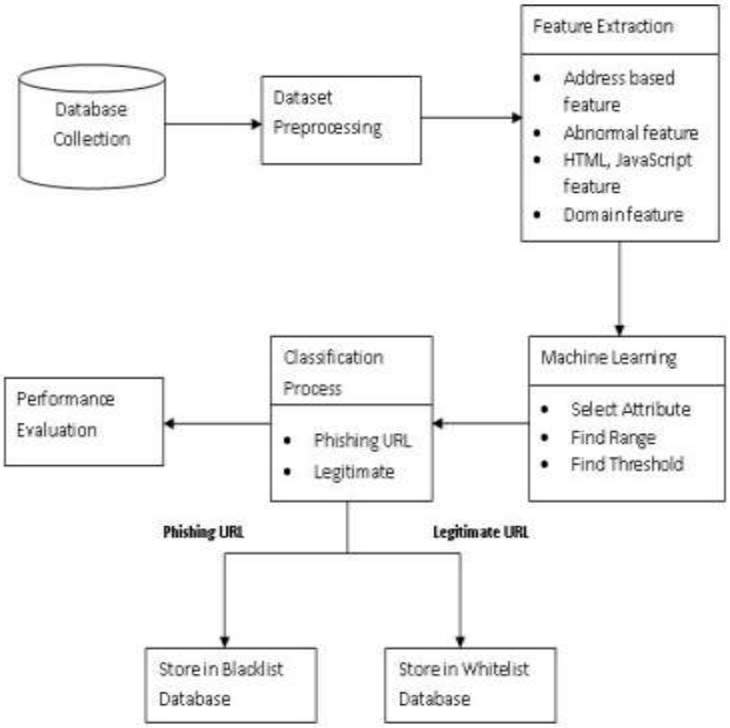

Figure 1 Proposed System

3.1 Modules Explanation

The proposed methodology imports data-set of phishing and legitimate URLs from the database and then the imported data is pre-processed. Identifying phishing website is performed based on following category of URL features: domain, address, abnormal and HTML, JavaScript features. The URL features are extracted with processed data and values for each URL attribute. The analysis of URL is performed using machine learning technique which computes the range value and the threshold value for URL attributes. Then it is differentiated as phishing and legitimate URL. The attribute values are computed using feature extraction of phishing websites and it is used to obtain the range value and threshold value. The value for each phishing attribute is ranging from {-1, 0, 1} these values are defined as low, medium and high. The classification of phishing and legitimate website is based on the values of attributes extracted using different types of phishing categories and a machine learning approach.

3.2 URL Feature Analysis

The phishing features are extracted for each URL to find whether the website is phishing or legitimate. The URL_of_Anchor tag attribute is selected to identify the overlap values. The overlap value is the summation of selected attribute value which is combined with other attributes.

3.3 Finding Attribute Values

The attribute value of the URL is computed using corresponding set of attribute values {- 1, 0, 1}. The attributes URL_of_Anchor tag and Prefix_Suffix also have inter linked value and that needs to be computed for finding range and threshold value.

3.4 Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) is a feed-forward artificial neural network (ANN) model and it has a single hidden layer. For the ANN in order to ensure a high-performing learning, parameters such as threshold value, weight and activation are requires function must have the appropriate values for the data system to be modeled. In gradient-based learning, all of these parameters are changed iteratively for appropriate values. Thus, they may be slow and may produce low-performing results due to the likelihood of getting stuck in local minima.

In ELM Learning Processes, the ANN that renews its parameters as gradient-based, input weights are randomly selected while output weights are analytically obtained. As an analytical learning process substantially reduces the solution time and the likelihood of error value which gets stuck in local minima, it also increases the performance ratio. In order to obtain the cells in the hidden layer of ELM, a linear function as well as non-linear (sigmoid, sinus, Gaussian), non-derivable or discrete activation functions can be used.

3.5 PROPOSED ALGORITHM

```
1. Import and Preprocess Dataset.
2. Extract the features of URL
3. Compute attribute values, if
            Attribute present value = 1
            Attribute absent value = -1
            Attribute not considered = 0
    3.1 Select attribute X and Y
    3.2 Compute equation for X and Y
4. Calculate threshold value for attribute X and Y
5.  Find range value.
6. Select Attribute to get threshold value.
7. Distinguish phishing and legitimate site using attribute value.
8. Compute Sensitivity and Specificity.
```

.

4. CONCLUSIONS

We defined features of phishing attack and thus proposed a model in order to classification of the phishing attacks. It consists of feature extraction from websites and classification section. In the feature extraction, we defined rules of phishing feature extraction and these rules have been used for obtaining features. Every user should also be trained not to blindly follow the links to websites where they have to enter their personal information. It is necessary to check the URL before entering the website.

5. ACKNOWLEDGEMENT

I am colossally gratifying to Dr. K.T.V Reddy, Principal, Sir Visvesvaraya Institute of Technology (SVIT), Nashik for inspiring and providing implausible backing and leadership too.As well as prolong obligation towards Prof. K. N. Shendge, HOD M.Tech (CSE) of Computer Engineering Department, Prof. D. S. Thosar, Assistant Professor and M.E Coordinator and Staff Member for their appreciated assistance and provision.

6. REFERENCES

[1]. [1] N. Abdelhamid, A. Ayesh, F. Thabtah, “Phishing detection based associative classification data mining,” Expert Systems with Applications, vol. 41(13), pp. 5948-5959, 2014.
[2] R. M. Mohammad, F. Thabtah, L. McCluskey, “Tutorial and critical analysis of phishing websites methods,” Computer Science Review, vol. 17, pp. 1-24, 2015.
[3] H. Huang, S. Zhong, J. Tan, “Browser-side countermeasures for deceptive phishing attack,” Fifth International Conference on Information Assurance and Security IAS'09, vol. 1, pp. 352-355, IEEE, 2009.
[4] R. M. Mohammad, F. Thabtah, L. McCluskey, “Predicting phishing websites based on self-structuring neural network,” Neural Computing and Applications, vol. 25(2), pp. 443-458, 2014.
[5] 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.
[6] 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.
[7] P. A. Barraclough, M. A. Hossain, M. A. Tahir, G. Sexton, N. Aslam, “Intelligent Phishing Detection and Protection Scheme for Online Transactions,” Expert Systems with Applications, vol. 40(11), pp. 4697- 4706, 2013
[8] H. H. Nguyen, D. T. “Nguyen, Machine learning based phishing web sites detection,” In AETA 2015: Recent Advances in Electrical Engineering and Related Sciences , pp. 123-131, Springer International Publishing, 2016.
[9] R. M. Mohammad, F. Thabtah, L. McCluskey, “Intelligent Rule-based Phishing Websites Classification, “ IET Information Security, vol. 8(3), pp. 153-160, 2014.
[10] V. S. Lakshmi, M. S. Vijaya, “Efficient prediction of Phishing Websites Using Supervised Learning Algorithms,” Procedia Engineering, vol. 30, pp. 798-805, 2012.
[11] J. James, L. Sandhya, C. Thomas, “Detection of Phishing URLs Using Machine Learning Techniques,” International Conference on Control Communication and Computing (ICCC), pp. 304-309, IEEE, 2013.
[12] M. Al-diabat, “Detection and Prediction of Phishing Websites using Classification Mining Techniques”, International Journal of Computer Applications, vol. 147(5), pp. 5-11, 2016.
[13] A. Hodzic, J. Kevric, A. Karadag, “Comparison of Machine Learning Techniques in Phishing Website Classification,” 2016.