An Efficient Machine Learning based Algorithm for Preventing Phishing Websites

Peravali Kavya
B. Tech Computer Science and Engineering, GITMA University, Visakhapatnam Andhra Pradesh.

Abstract

Now a day's Internet technology is growing more pervasive for online technologies. By considering we cannot control the security of the application. Because of this, we can face security threats of the network system that is mostly encountered is Phishing. Phishing is one of the most attacked of web-based which try to reveal some sensitive information. So many people or companies information is attacked by attackers by using these techniques. To prevent phishing damages we can provide the most secure networks or to get awareness of the people. To detect or preventing the phishing attacks we can build strong mechanism before they cause too much damage. In this paper, we are proposed an efficient machine learning based detection schema to detect or prevent the phishing attack. By implementing this technique we can get zero hour phishing attacks and they have superior adaption for new types of phishing attacks, therefore they are mainly preferred.

Keywords: *Phishing Attack, Entropy, Gain, Machine learning, URL, Domain Names.*

I. INTRODUCTION

Phishing is a type of extensive fraud that happens when a malicious website acts like a real one keeping in mind that the end goal to obtain touchy data. In a Web phishing attack, phishing websites are created by the attacker, which are similar to the legitimate websites to deceive Web users in order to obtain they're sensitive financial and personal information. The phishing attack is initially performed through clicking a link received within emails. Victims receive an email containing a link to update or validate their information. If this link is clicked by the target victims, the Web browser will redirect them to a phishing website that appears similar to the original website. The attackers can then steal the important information of the web users since they are asked to input the sensitive information on the phishing website.

Eventually, the attackers can carry out financial theft after phishing occurs [1]-[2]. Due to the inevitability of phishing websites targeting online businesses, banks, Web users, and government, it is essential to prevent Web phishing attacks in the early stages. However, detection of a phishing website is a challenging task, due to the many innovative methods used by phishing attackers to deceive web users [3]-[4]. The success of phishing website detection techniques mainly depends on recognizing phishing websites accurately and within an acceptable timescale [7]. Many conventional techniques based on fixed black and whitelisting databases have been suggested to detect phishing websites. However, these techniques are not efficient enough, since a new website can be launched within few seconds. Therefore, most of these techniques are not able to make an accurate decision dynamically on whether the new website is phishing or not.

Hence, many new phishing websites may be classified as legitimate websites [6], [7]. As alternative solutions to the conventional phishing website detection techniques, some intelligent phishing detection methods have been developed and suggested in order to effectively predict phishing websites. In recent years, the intelligent phishing website detection solutions based on supervised machine learning techniques have become common, which are smart and more adaptive to the Web environment compared to the conventional phishing website detection methods. As alternative solutions to the conventional phishing website detection techniques, some intelligent phishing detection methods have been developed and suggested in order to effectively predict phishing websites. In recent years, the intelligent phishing website detection solutions based on supervised machine learning techniques have become common, which are smart and more adaptive to the Web environment compared to the conventional phishing website detection methods.

II. RELATED WORK

The Phishing site is one of the late worries in the security area. Yet, because of prominent effect on the money related and online retailing areas and since recognizing such sort of dangers is key towards safe web surfing, various distinctive a few promising studies and methodologies were led and proposed to this issue established in the writing. Although a considerable amount of hostile to phishing arrangements are accessible these days yet a large portion of them are not skilled to settle on a

sufficiently precise choice and thus, the false-positive choices raised seriously.

In this segment, we quickly depict the existed endeavours in this space by reviewing the basic related methodologies. Nawafleh and Hadi[8] proposed a new associative classification algorithm to recognizing phishing site. Observational study result demonstrates that acquainted classification is a promising technique and indicated competitive execution when contrasted and different calculations, for example, SVM, PRISM, RIPPER and NB. In [9] the study compared a few learning approaches including Support-Vector-Machine, decision-trees, rule-based techniques and Bayeph phishing techniques in recognizing phishing emails. A random forest algorithm was executed in PILFER (Phishing Identification by Learning on Features of Email Received) which succeeded in effectively identifying 96% of the phishing messages with a false-positive rate of 0.1%. Ten email's elements showed 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, Number of Domains, Number of Dots, Containing JavaScript, Spam-channel Output. With respect to phishing location, A. Bergholz et el. [10] exhibited a methodology for enhancing learning models for recognizing phishing messages by feature selection. A subset of components is chosen by a wrapper technique in which the purported best-first pursuit calculation efficiently adds and subtracts features to a present subset utilizing the classifier itself as a feature of the evaluation function.

III.PROPOSED SYSTEM

Phishing is one of the web page based attach, that could be considered as url based document classification. By classifying document we can get document URLs with domain names. By taking those URLs we can detect the web page malicious or not. To Prevent or detecting malicious URLs we can use machine learning based algorithms are required. By implementing machine learning algorithms we can get zero hour phishing attacks. In this paper, we are proposing the Random forest decision tree based preventing algorithm for detecting or preventing malicious URLs. By implementing this algorithm we can get an efficient or zero hour phishing attacks.

Detecting phishing is one of the classification problems of data mining. By implementing a classification approach we can analyse data and gets attacked urls from the dataset. Therefore labelled data are necessary for performing classification process. By performing data classification we can take domain related samples like phishing domain or legitimate domain in the training phase are needed. In the training phase data set is used one of the crucial points to build successful detection of phishing attacks. The detection of phishing is considered to use whose classes are precisely known. So, the samples which are labelled as phishing must be absolutely detected as phish. Likewise, the samples which are labelled as legitimate must be absolutely detected as legitimate. Otherwise, the system cannot work correctly if we use samples that we are not sure about the class information. For this purpose, a number of public datasets are created for phishing

Another problem of classification is collecting legitimate domain, for this purpose we are used site reputation services. The site reputation service provides a ranking to all available websites. The ranking of these web sites is providing globally or may be country based ranking. To provide ranking of web sites we can consider various features depends on the mechanism. The websites which have high rank scores are identified as legitimate sites which are used very frequently. To perform the phishing we have to use raw data for processing these data and getting meaningful information from it to detect fraudulent domains. In this paper, we are taking machine learning dataset which consists of labelled related features. By taking raw data and create a training dataset with machine learning algorithms. The values should be selected according to our needs and purposes and should be calculated for every one of them.

To Build a Phishing detection mechanism system should calculate features that we have selected to our needs and purpose with labels. By implementing the following algorithm we can detect or preventing urls using labels. The Implementation of Random forest decision tree based preventing algorithm is as follows.

1. Read training data from the data set with contains labels of Protocol, Domain Name, URL Length, Query String and Classification Status.

2. After completion of the reading process we can construct a decision tree by implementing the following process.

i. Choose the root node for the tree in given training data.

ii. If the training data set(S) contains positive then the leaf nodes are positive.

iii. If the training data set contains negative then the lead nodes are negative.

iv. By taking training data set we can calculate entropy H(S) of every attribute an of the data set S.

V. Entropy H(S) is a measure of the amount uncertainty in the data set S.

Vi. We can calculate entropy by using the following formula.

$H(S) = - \sum p(x) \log_2 p(x)$

S is the current dataset which entropy is calculated
X is the set of classes in the dataset S.
P(x) is the proposition of the number of elements in class x to the number of elements in set S.

Vii. Calculate each attribute entropy with respect x denoted by H(S, x).

Viii. After completion of calculation, we can calculate smallest entropy is used to split set S in to set of iterations. The information theory of entropy measures how much information is expected to be gained upon a measuring random variable.

ix. The calculated entropy is used to quantify amount to which the distribution of the quantity's values is unknown. A constant quantity has zero entropy, as its distribution is perfectly known. In contrast, a uniformly distributed random variable (discretely or continuously uniform) maximizes entropy. Therefore, the greater the entropy at a node, the less information is known about the classification of data at this stage of the tree; and therefore, the greater the potential to improve the classification here.

X. The range of entropy is 0 perfectly classified and 1 is totally random.

Xi. After completion of calculating the entropy of each attribute, we can find out the gain of each attribute in a training data.

Xii. The calculation of gain can be done by using the following equation.

$Gain(S,A) = Entropy(S) - \sum (|S_v|/|S|) * entropy(S_v)$

Where an S_v= subset of which attribute A has values V

S_v = number of elements in S_v
S=number of elements in S

$\sum$ is each value v of all possible value of attributes A

3. After calculating each attribute gain values and find out the highest of the value of attributes.

4. Take that attribute is a decision root node and this process run goes until all data is classified perfectly or we run out of attributes.

5. After completion of this process, we can get randomized decision tree with contains classified data with attributes.

6. The completion of tree construction take the testing data who's URL are preventing or phishing.

7. By taking each record from the testing data set and apply the phishing process on the tree.

8. By performing the searching process of the tree we can get the status result of particular record based on the tree.

9. Take that result as phishing of URL and apply this process until completion of testing data.

By implementing this process we can get an efficient phishing result of url and also get non zero hours based to prevent the attacks.

IV. CONCLUSIONS

In this paper, we have proposed an efficient phishing detection or preventing of URLs with zero hour attacks. Now a day's phishing of attacks is a major problem of internet technology, which uses both social engineering and technical deception. By performing phishing attacks to get users important information such as financial data, emails, and other private information. Phishing exploits human vulnerabilities; therefore, most protection protocols cannot prevent the whole phishing attacks. Many of them use the blacklist/whitelist approach, however, this cannot detect zero-hour phishing attacks, and they are not able to detect new types of phishing attacks. By preventing those phishing attacks we can implement Random forest decision tree based preventing algorithm. By implementing this process we can get non zero hour phishing attacks of machine learning training data set.

REFERENCES

[1] 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.
[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, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25(2), pp. 443-458, 2014.

[4] 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.

[5] 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.

[6] N.Abdelhamid, A. Ayesh, F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41(13), pp. 5948-5959, 2014.

[7] R.M.Mohammad, F. Thabtah, L. McCluskey, "Tutorial and critical analysis of phishing websites methods," Computer Science Review, vol. 17, pp. 1-24, 2015.

[8] S.Nawafleh, W. Hadi (2012). Multi-class associative classification to predicting phishing websites. International Journal of Academic Research Part A; 2012;4(6), 302-306J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[9] Sadeh N, Tomasic A, Fette I. Learning to detect phishing emails. Proceedings of the 16th international conference on World Wide Web. 2007: p. 649-656.

[10] Andr Bergholz, Gerhard Paa, Frank Reichartz, Siehyun Strobel, and Schlo Birlinghoven. Improved phishing detection using model-based features. In Fifth Conference on Email and Anti-Spam, CEAS, 2008

[11] S.Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of the 6th Conference in Email and Anti-Spam, ser. CEAS'09, Mountain view, CA, July 2009.

[12] P.Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in INFOCOM'10: Proceedings of the 29th conference on Information communications. Piscataway, NJ, USA: IEEE Press, 2010, pp. 346–350.