

Phishing Identification Using An Efficient Neuro-Fuzzy Model

Ms. Roshni Vitthal Pawar¹

Department of Computer Engg.,
K.K.Wagh Institute of Engineering Education and Research
Nashik, India

Ms. Ruchita Panditrao Pawar²

Department of Computer Engg.,
K.K.Wagh Institute of Engineering Education and Research
Nashik, India

Ms. Pranali Ganesh Salunkhe³

Department of Computer Engg.,
K.K.Wagh Institute of Engineering Education and Research
Nashik, India

Ms. Ankita Gajanan Sankhe⁴

Department of Computer Engg.,
K.K.Wagh Institute of Engineering Education and Research
Nashik, India

Abstract: *Now a days people use internet on their daily basis. It has many benefits and potential risks too. In today's world, cyber-crimes like hacking are growing rapidly, phishing is one of the new type of online crime. Phishing site is a fake site aimed to steal personal information such as banking account and credit card information etc. Most of these phishing pages look similar to the real pages in form of interface and uniform resource locator (URL) address. However, the numbers of victims have been increasing due to inefficient protection technique. In this paper, we develop a neuro-fuzzy model for phishing identification efficiently. The model eliminates the subjective factors to improve efficiency such as if-then rule sets, the parameters of membership functions, etc.*

Keywords: *Phishing, URL-Based, Neuro-Fuzzy.*

I. INTRODUCTION

Phisher use a various techniques to fool their victims aimed to steal the personal information including email messages, phone calls and social networking. These activities of phishing cause severe economic loss all over the world. Phishing sites are also growing rapidly in quality and quantity. Therefore, the risk of stealing user information is extremely high. For of these reasons, identifying phishing problem is complex and extremely important problem.

In this paper, an efficient method is proposed to identify the phishing sites that focus on the features of URL like Primary-Domain, Sub Domain, Path Domain and Google's parameters PageRank, Back Link, Google Index. Then, a proposed neuro fuzzy model is a system which reduces the error and increases the performance. The neuro-fuzzy model uses computational models to perform without using if-then rule sets.

II. LITERATURE SURVEY

There are many methods for identifying phishing. These can be divided into three groups: blacklist, heuristic and machine learning. The blacklist -based technique [2][3][4][9] maintains a list of phishing websites called blacklist. The technique is inefficient due to the rapid growth in the number of phishing sites. Therefore, the heuristic and machine learning approaches have received more attraction of researchers.

In the rule based phishing technique [5], in which the phishing websites are detected using if-then rule sets. It is the process of detecting the unseen knowledge from the dataset is represented in the terms of rules and it is known as rule-induction. It predicts the phishing websites using rules derived from different rule induction algorithm. This technique uses rule-based systems which are costly and time consuming.

The fuzzy technique is based on 27 features of webpage, classified into 3 layers. Each feature has three linguistic values: low, moderate, high. The technique has built a rule set, triangular and trapezoidal membership functions. The achieved rate of the technique is

86.2%. But, there exist many drawbacks. First, the rule sets are not objective and greatly depend on the builder. Second, the weight of each main criteria is used without any clarification. Finally, the used heuristics are not optimal and really effective.

In the neural network technique[8]. Three layers were used in the neural network including input layer, hidden layer and output layer. The best achieved rate of the technique is 95%. However, there exist some drawbacks. First, a number of hidden nodes and activation function must be determined through experimentation. Second, the authors do not explain why using one hidden layer. Third, the value of features does not know how it is calculated. Finally, the datasets are not enough to verify the accuracy. With respect to previous techniques, URL plays a minor role for identifying phishing websites.

In this paper, we design a new neuro-fuzzy model based on URL's features and Google's parameters to identify phishing sites. The model without defining the if-then rule sets is developed from with new aspects: i)The new heuristics have been proposed to identify phishing website more effectively and rapidly. ii) The parameters of the membership functions are eliminated, so the fuzzy values are calculated more objective. iii) The input values are normalized to enhance the accuracy and the convergence of training phase. Besides, The other aspects also supports the new model as follows: i) The weights are trained by neural network, so the model is more efficient. ii) The if-then rules are not utilized. Hence, the result will be more precise and objective.

III SYSTEM DESIGN

- System Model Design:

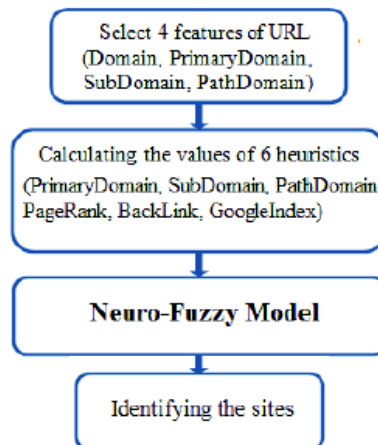


Fig.1.System Model

A. 1.URL and Its Features:

A URL (uniform resource locator) is used to locate the resources. The structure of URL is as follow:
 $\langle protocol \rangle : // \langle subdomain \rangle . \langle primarydomain \rangle . \langle TLD \rangle / \langle pathdomain \rangle$

B. 2.Six Heuristic Values Of URL

The six heuristic values of URL are as follow: Primary Domain , Sub Domain, Path Domain, Pagerank, Backlink , Google index.

C. 3.Neuro-fuzzy Network Model-

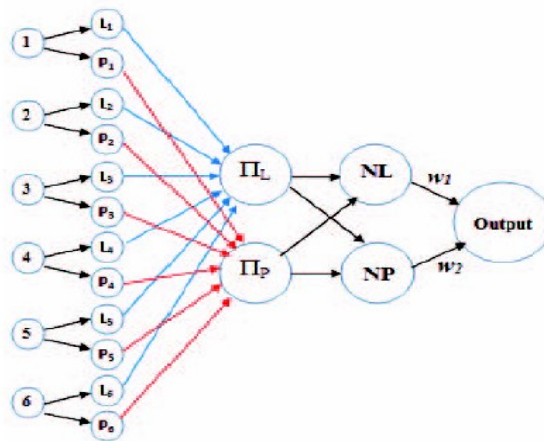


Fig. 2. The neuro-fuzzy network model

1) The neuro-fuzzy network model was designed with five layers as follows:

- The first layer, called the input layer, contains six nodes that are six heuristics such as PrimaryDomain, SubDomain, PathDomain, PageRank, BackLink, GoogleIndex.
- The second layer contains 12 nodes. The value of each node is calculated from the left sigmoid membership functions and the right sigmoid membership function.
- The third layer contains two nodes these are as follows:

$$\pi_L = \prod_{i=1}^6 L_i$$

$$\pi_P = \prod_{i=1}^6 P_i$$

- The fourth layer contains two nodes which are NL (Normalization Legitimate) and NP (Normalization Phishing).

$$NL = \frac{\pi_L}{\pi_L + \pi_P}$$

$$NP = \frac{\pi_P}{\pi_L + \pi_P}$$

- The fifth layer, called the output layer, has one output node. The output value of the output node ranges from 0 to 1. The site is phishing if the value of the output node is less than 0.5 and the site is legitimate, if the value is greater than or equal to 0.5.

2) The value of input node is converted into new values using the following formula .Where, Max is 10 and Min is -10. $Value_{old}$ ranges from 0 to 1. $Value_{new}$ ranges from -10 to 10.

$$Value_{new} = Value_{old} * (Max - Min) + Min$$

3) The value of 12 nodes in the second layer is calculated using the following formula,

$$L(x) = \frac{1}{1 + e^{-x}}$$

$$P(x) = \frac{e^{-x}}{1 + e^{-x}}$$

4) For the output node, input value of output node and output value of output node are calculated by,

$$O_I = \sum_i W_i * I_i$$

$$O_O = \frac{1}{1 + e^{-O_I}}$$

Where O_I and O_O are input value of output node and output value of output node respectively.

IV. EXPERIMENTAL SETUP

- Hardware
 - 1.Processors : Pentium IV.
 - 2.HDD : 20 GB Min 40 GB Recommended
 - 3.RAM : 1 GB Min 2 GB Recommended
 - 4.Keyboard : Standard 102 keys
 - 5.Mouse : 3 buttonn
- Software
 1. Operating System: Windows
 2. IDE : Eclipse
 3. Database : My-SQL
 4. Technologies : JavaScript
- Dataset
It contain two data sets those contain phishing sites and legitimate sites.

CONCLUSION

We have proposed a new technique to identify phishing sites efficiently. In the technique, the system model is built to identify phishing sites by using neuro-fuzzy network with five layers and six heuristics (PrimaryDomain, SubDomain, Path-Domain, PageRank, BackLink, GoogleIndex). The best accuracy result can be obtained.

REFERENCES

1. Phishing Identification: An Efficient Neuro-Fuzzy Model Without Using Rule Sets (2015)
2. PhishTank. (2013, Nov.) [Online]. Available: <http://www.phishtank.com>
3. D. Goodin. (2012) Google bots detect 9,500 new malicious websites every day. [Online]. Available: <http://arstechnica.com/security/2012/06/>
4. S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang. (2009) An empirical analysis of phishing blacklists. [Online]. Available: <http://ceas.cc/2009/papers/ceas2009-paper-32.pdf>
5. Intelligent rule based phishing identification website classification.
6. Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in The 16th international conference on World Wide Web, 2007, pp. 639—648
7. G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: a feature-rich machine learning framework for detecting phishing web sites," *ACM Transactions on Information and System Security*, vol.14, no.2 .pp. 1—28, Sept. 2011.
8. M. Aburrous, M. Hossain, F. Thabatah, and K. Dahal, "Intelligent phishing website detection system using fuzzy techniques," in Third International Conference on Information and Communication Technologies: From Theory to Applications, 2008, pp. 1—6.
9. McAfee. (2011, July) McAfee site advisor. [Online]. Available: <http://www.siteadvisor.com>
10. N.Zhang and Y. Yuan, "PhishingDetectionUsingNeuralNetwork", CS229 lecture notes, <http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf>, 2012