The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 5 || Pages || PP -103-106 || 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Phishing Websites Detection Using Back Propagation Algorithm: A Review

[1]Amandeep Sharma, [1]Pardeep Singh, [2]Amrit Kaur
[1]*M.tech (CE) Research Scholar, Department of Computer Engineering, Punjabi University, Patiala, India*
[2]*Assistant Professor, Department of Computer Engineering, Punjabi University, Patiala, India*

ABSTRACT
Phishing is an illicit modus operandi employing both societal engineering and technological subterfuge to theft client's private identity data and monetary account credentials. Influence of phishing is pretty radical as it engrosses the menace of identity larceny and financial losses. This paper elucidates the back propagation paradigm to instruct the neural network for phishing forecast. We execute the root-cause analysis of phishing and incentive for phishing. This analysis is intended at serving developers the effectiveness of neural networks in data mining and provides the grounds proving neural networks in phishing detection.

Date of Submission: 09 May 2016 Date of Accepted: 23 May 2016

I. Introduction

Data mining is the non-trivial withdrawal of valuable predictive information from enormous databases. The type and size of the data stored mainly depends on the corporation. For the effectual data mining, diverse things required are: superior quality data, precise choice of data, perfect sample size and ample data mining tool. Advantages of data mining have amplified its applications to various fields such as medical, banking and phishing website detection etc.

With the boost in technology, tradition of internet too increases. Most of the services similar to banking, shopping etc are online nowadays. This elevates the attacks in cyberspace. Phishing is a mode of deception which is planned to ploy the individual to reveal their sensitive information like username, password etc. This information is used by the phishers to filch the identities of the victim or for illegal financial gains.
Fig.1 indicates the unique phishing websites detected since the year 2007 to 2015.

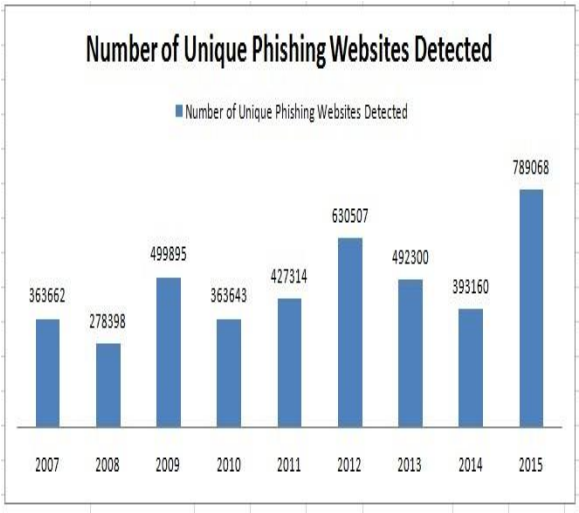

Fig. 1

APWG (anti phishing working group) was published many phishing attack trends reports on phishing [1]. These reports inform about often changes in the phishing attacks detected in the last few years.
A variety of data mining tools and techniques are available to spot the phishing websites. But it is complicated to pick most adequate technique or tool for the researchers.

Our review emphasises on effectiveness of neural networks in predicting phishing websites and grounds behind using the neural networks to detect them. As a consequence we deliberate various articles and research papers which notify the increased curiosity of researchers in this area.

Phishing motives:

Weider.D.Et.al (2008)[3] portrays an assortment of phishing vulnerability intention of the phisher as follows:

1. **Financial gain:** Attacker stoles the banking credentials for their financial gain.
2. **Identity hiding**: identities of the victim can be used via attackers to attain merchandise and services and occasionally they are sold to other individuals who might be a criminal.
3. **Fame and notoriety:** sometimes phishers mainly attack for the fame and popularity in the social media.

Necessity of phishing detection

Phishing is the massive problem in today`s scenario as it is effecting both consumers and vendors. APWG (Anti Phishing Working Group) is an anti-phishing cluster which published plenty of reports on phishing. The unique phishing websites detected during the year 2012 be 630507 [1]. It decreases in the year 2013 and 2014 as compared to year 2012 as the attacker disappears. The decline in the attacks is due to the switching in the actions of phishers from conventional phishing to malware phishing.

According to APWG report, there is a immense increase in the unique phishing websites detected in the year 2015. There are 789068 unique phishing websites detected in this year [1]. Various organisations and end users are effected by phishing. These attacks are not restricted to the naivety of end users. Technological engineers can also be fatalities of these attacks. ”Business email compromise” (or BEC) cons turn out to be a chief trouble in 2015. These attacks lead to huge financial and identity loses to organisation and consumers.
Therefore, it is extremely vital to diminish the shock of the phishing attacks.

Objective:

The main objective of this review is to explore the diverse aspects of the phishing. We uncover the utility of data mining to detect phishing websites by surveying the literature. Effectiveness of neural networks and training of neural networks through back propagation algorithm to predict phishing is also discussed.

Related Work:

Various researchers investigate the phishing websites and proposed various tools and techniques to detect them. Nguyen (2014) [6] presents an efficient approach for phishing detection using single layer neural networks. The results of the proposed technique are evaluated with the help of dataset of 11,660 phishing and 10,000 legitimate websites. This technique evaluates the weights of heuristics using single layer neural networks and the precision of the system is up to98%.

Zhang et.al (2007) [12] projected a content based scheme CANTINA which is a linear trait classifier. Developed scheme is based on TF-IDF algorithm to condense the false positives.
In year 2011, Xiang et.al [13] improved the technique CANTINA as CANTINA+ by using two dissimilar gamuts of corpora. Two filters are employed to dilute the false positives and to consummate the runtime speedup.

Saeedabu-nimehet.al (2007)[10] contrasts various machine learning techniques for phishing websites detection. They formulate the comparison of machine learning techniques in predictive phishing including classification and regression trees, support vector machines, neural networks, logistic regression, Bayesian additive regression trees and random forests. They used 2889 phishing and legitimate emails for their study and 43 supplementary features to test and train the classifiers.

Khonji et.al (2013) [11] surveys the literature for the detection of phishing attacks. Various phishing alleviation techniques are deliberated. This research presents many categories of phishing mitigation techniques like: detection, offensive defence, correction and prevention. A.martin et.al (2011) [5] develops a framework for predicting the phishing websites. They used the neural networks to predict the phishing websites. According to their study, multilayer neural networks shrink the error and elevate the performance.
Krutika rani sahu and jigyasu dubey (2014) [8] conducted a survey on phishing attacks. In their survey, they discover various aspects regarding the phishing attacks, their problems and presented the problem statement for searching the finest solution for the problem. They also provide a structural design of a system intended for future simulation of internet security based security.

Neural Networks within Data Mining:

An Artificial Neural Networks are defined as the computational model with meticulous properties like the capability to learn or adopt, to simplify, to categorize or to systematize data and which operation is based on parallel processing [7]. In realistic terms, the neural networks are statistical data modelling tools which are non-linear in nature. Neural networks have convoluted relationships which are modelled between inputs and outputs

to uncover the patterns in the data. Data warehouse firms employ the neural networks as tool to pull out information from datasets in the process called data mining.

A structural design, learning paradigm, and activation function are the three pieces enclosed in a neural network. Neural networks are trained to accumulate, recognize and associatively regain patterns to crack the optimization problems. Efficacy of neural networks in data mining is owing to the pattern recognition and function evaluation properties. Due to "model free" estimators and dual nature, neural networks constitute data mining in countless ways [2].

Classification is the widespread action in data mining. It recognizes patterns which convey regarding the cluster to which an item suits. It inspects the existing items which are already classified and deduce a set of rules to categorize the items. Neural networks harvest precious information from large history information on origin of which items are classified and to infer rules for them.

Grounds proving Neural Networks

In neural networks, classification is the most investigated issue. Zhang [4] in his research presented various vital issues and advancements in neural networks which comprise posterior probability estimation, the association between neural and conventional classifiers, the bond between learning and generalization in neural networks classification and issues to enhance neural classifier performance. His research provides that neural networks are competitive alternative to traditional classifiers for many realistic classification problems.

Artificial neural networks are encouraged from the biological systems. It consists of extremely uncomplicated but large amount of nerve cells that work parallel and have the facility to learn. Explicit program is not needed for neural network because it can be trained from training samples (reinforcement learning).

Neural networks have the ability to generalize and correlate data. Subsequent to training, an artificial neural network can find the sensible solutions for the same sort of problems of the identical class that were not explicitly trained which results in the high scale of fault tolerance for noisy input data [9].

Phishing Websites Detection using Back Propagation Algorithm

Neural networks have broad range of applications such as financial forecasting, phishing detection etc. Phishing detection is a classification problem. Owing learning & generalization feature, neural networks are used to forecast phishing. Neural network works in the similar way of human being processes information. These are the computational or mathematical models which are used to locate the solution for various combinatorial optimization problems. In neural networks, copious neurons work concurrently to solve the problem. Neural networks require knowledge from a learning process while the interneuron association potency called synaptic weights accumulates knowledge. The values of the heuristics of the website are given as the input to the input layer of the neural network.

The learning feature of neural networks makes it suitable solution for classification or pattern recognition problems. Neural networks guide when examples with acknowledged results are given to it. The weights are attuned by an algorithm to convey the final result nearer to the recognized result. Neural networks are feed forward networks. A feed forward network is instructed via back propagation algorithm.

A feed forward network has layered structure. A simplest feed forward network encloses three layers: input layer, hidden layer and output layer. One or more processing elements are there in each layer. Each processing unit gets input from the outside world or the previous layer. Processing elements are allied to each other and having weights associated with them. In these networks, information pours in forward direction throughout the network. No feedback cycles are present in these networks.

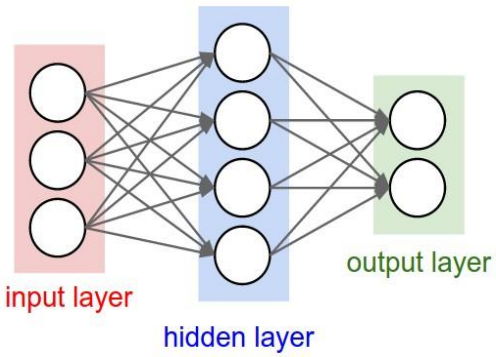

Fig 2: Simple Neural Network

The simple training process (Back Propagation Algorithm) of neural networks to detect phishing websites is [2][6] :

1. Required features of the website are selected and then the weights are intended for the heuristics.
2. These weights are offered to the network as the input and broadcast throughout the network in forward direction until it arrives at the output layer. The input for output layer is calculated by equation:

$$O_I = \sum_{i=1}^{n} W_i * I_i$$

Where O_I is the input value for the output layer, I_i is the value of the ith input node and W_i is the weight of the ith input node. The equation used to calculate the output value of output node is:

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

Where O_O is output value of output node and O_I is the input value of output node. Output of this system is legitimate, suspicious or phishing.

3. The output value of the output layer is deducted from the real output value to calculate the error.

$$err = T - O_O$$

Where T is the actual output.

4. Supervised learning is utilized to train the neural networks. Back propagation is used in this case. In the back propagation learning algorithm, weights are adjusted by the equation:

$$W_i = W_i + R * err * O_O$$

Where R is the learning rate and W_i is the weight of the ith input node. This algorithm commences with weights between output layer processing elements and the hidden layer and then works towards the back through the network.

5. When the back propagation has stopped, the forward process commences again, and this cycle persists until the error between predicted output and genuine output is diminished.

II. Conclusion

Phishing is a means of deception in which an individual is ruse to reveal their credentials similar to user name and passwords. It is vital to forecast the phishing websites as it origins the identity larceny and financial fatalities. Phishing is a classification problem. Consequently, neural network is one of the preeminentsolutions to this problem as they owing the learning and pattern recognition properties. They are fault tolerant in nature. When any neuron of the neural network aborts, it can persist devoid of any dilemmaon account of its coextending nature. Back propagation is the simple algorithm to guide the neural networks. We deemed that neural network works well and confersaninferior error rate.

REFERENCES

[1] Anti Phishing Working Group. (2007-2016). *APWG Phishing Attack Trends Reports.* APWG.

[2] Dr. Yashpal Singh and Alok Singh Chauhan. (2005-2009). Neural Networks In Data Mining. *Journal of Theoretical and Applied Information Technology* , 37-42.

[3] W. D. Yu, S. Nargundkar, and N. Tiruthani. (2008). A Phishing Vulnerability Analysis ofWeb Based Systems. *13th IEEE Symposium on Computers and Communications (ISCC 2008).* (pp. 326-331). Marrakech,Morocco:: IEEE.

[4] Zhang, G. P. (2000, november). Neural Networks for Classification: A Survey. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, ,* 451-462.

[5] A.Martin et.al. (2011). A Framework For Predicting Phishing Websites Using Neural Networks. *International Journal of Computer Science Issues, 8* (2), 330-336.

[6] Nguyen et.al. (2014). An Efficient Approach for Phishing Detection Using Single Layer Neural Network. *The 2014 International Conference on Advanced Technologies for Communications* (pp. 435-440). hanoi: IEEE.

[7] Ben Kr ose, P atrick van der Smagt. (1996). *An Introduction to Neural Networks* (8th ed.). Amsterdam: The University of Amsterdam

[8] Krutika Rani Sahu, Jigyasu Dubey. (2014). A Survey on Phishing Attacks. *International Journal of Computer Applications, 88*, 42-45.

[9] Kriesel, D. (2005). *A Brief Introduction to Neural Networks.* www.dkriesel.com .

[10] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang and Suku Nair. (2007). A Comparison of Machine Learning Techniques for Phishing Detection . *eCrime '07 Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit* (pp. 60-69). New York, USA: ACM.

[11] Mahmoud Khonji ; Youssef Iraqi ; Andrew Jones. (2013). Phishing Detection: A Literature Survey. *IEEE Communications Surveys & Tutorials, 15* (4), 2091-2121.

[12] Yue Zhang, Jason Hong, Lorrie Cranor. (2007). Cantina: a content-based approach to detecting phishing web sites. *WWW '07 Proceedings of the 16th international conference on World Wide Web* (pp. 639-648). NEW YORK, USA: ACM.

[13] Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor. (2011). CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Transactions on Information and System Security (TISSEC), 14* (2), 1-28.