Australian Journal of Basic and Applied Sciences, 10(12) July 2016, Pages: 213-219

AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

Hybrid Intelligent Approach in Prediction of Phishing Attacks

[1]Sarala R, [2]Harshini E, [3]Nandini S, [4]Zayaraz G

[1]Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry-14,India.
[2]Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry-14,India.
[3]Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry-14,India.
[4]Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry-14,India.

Address For Correspondence:
Sarala R, Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry- 14, India.
E-mail: sarala@pec.edu

ARTICLE INFO

Article history:
Received 26 April 2016
Accepted 21 July 2016
Published 30 July 2016

Keywords:
Phishing attacks, Hybrid approach, Feature set reduction, Ant colony optimization, Artificial neural network.

ABSTRACT

Phishing website is a fake website that mimics the appearance of actual website but leads to a different destination. The users provide their data thinking that these websites come from genuine financial institutions. Phishing is an unceasing problem where features significant in determining the types of web pages are constantly changing. In this paper, a hybrid intelligent heuristic method is proposed to determine whether a webpage is legitimate or phishy**.** The proposed work aims to identify the most significant features using the ant colony optimization, bio inspired algorithm and build an artificial neural network adaptively to detect the phishing websites.

INTRODUCTION

Internet is extremely important for individual users and also for different organizations for sharing information. Internet users may be vulnerable to many web threats that may cause financial damage, loss of private information and loss of customer's confidence in electronic commerce and online banking. Phishing is considered a form of web threat that is defined as the art of imitating a website of an trusted enterprise aiming to acquire personal information (Ramzan.S,2010). These involve malicious websites that look authentic; they may even appear to be the real website of the company people looking for, though they are fake.

The goal of a phishing swindler is to get access to any information such as user credentials, which can be collected via fraudulent websites. Phishing attacks can be of many types such as spear phishing, clone phishing, whaling (Ramzan.S,2010). Phishing attacks can lead to personal risks like using data to access victim's account and withdraw money or can open bank accounts in a victim's names, and use the new account to cash illegitimate checks. It can also lead to institutional risks such as potentially accessing high value institutional data resulting in reputational damage.

Anti-phishing measures (Mohammad.R.M. *et al*.,2014) that are implemented take several forms including legal, education and technical solutions. The technical solutions include:

- Blacklist approach: In this approach, a blacklist containing url's of phishing websites are predefined. The requested url is compared with blacklist to identify whether it is phishing or legitimate. The drawback of this approach is the freshly created fraudulent websites cannot be instantly updated in the blacklist.
- Heuristic approach: It is also known as feature- based methods, where several features are extracted from the websites to classify it as phishing or legitimate. When compared to the blacklist approach, it can recognize newly created phishing websites in real time. A heuristic

Open Access Journal
Published BY AENSI Publication
© 2016 AENSI Publisher All rights reserved
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/

Open Access

To Cite This Article: Sarala R, Harshini E, Nandini S, Zayaraz G., Hybrid Intelligent Approach in Prediction of Phishing Attacks. ***Aust. J. Basic & Appl. Sci.,*** *10(12): 213-219, 2016*

based approach is used in the proposed system where ACO is combined with artificial neural networks to determine the phishing websites.

Artificial Neural Networks:

Artificial neural networks are the group of models inspired by biological neural networks which are composed of interconnected processing units called neurons (McCulloch.W.S. and Pitts.W., 1943). The connections have numeric weights that can be tuned based on experience, making neural network adaptive to inputs and to learn. ANN's is particularly used in the proposed system as they have the following advantages:

- require less formal statistical training to develop.
- can detect complex nonlinear relationships between independent and dependent variables.
- can be developed using multiple training algorithms.

ANNs are used and proved to be successful in function approximation, regression analysis, data processing, robotics and classification problems(Ganesan.N et.al.,2010;Hashimoto.H *et al*.,1992). Phishing detection comes under classification and hence ANNs can be used for the prediction of phishing attacks. The efficiency of output from artificial neural network depends on its architecture and so to reduce the complexity in the architecture, the number of neurons is controlled by employing a feature selection technique.

Ant Colony Optimization:

The main objective of feature selection is to determine a minimal feature subset from a problem domain while maintaining high accuracy in representing the original features. In feature selection problems, some features contain relevant information about output, whereas the other features contain less information regarding output. The task for feature selection is to find the input features that contain more information about the output. As feature selection is NP-hard problem, bio inspired algorithms such as ant colony optimization (ACO) (Socha.K and Dorigo.M.,2008), particle swarm optimization (Das.G. *et al*.,2010), simulated annealing (Eglese.R.W.,1990) can be used.

Ant colony optimization is used to derive an optimal feature set from the large set. It is a metaheuristic approach in which a group of artificial ants work together in finding good solutions to discrete optimization problems. The most interesting aspect of the collective behavior of various ants lies in their ability to find minimal paths between the ants' nest and food sources by tracing the pheromone deposits. Then the ants choose to follow the path by the amount of pheromone deposited: the stronger the pheromone, the higher its desirability. ACO requires a problem to be depicted as a complete graph where nodes denote features and the edges between them denote the choices of next features. The advantage of this algorithm over simulated annealing and genetic approach is that when the graph changes dynamically, it can run continuously and can adapt to changes in real time.

The rest of this paper is organized as follows. Section2 gives an overview of work related to prediction of phishing attacks and bio inspired algorithms. The proposed hybrid intelligent approach is discussed in Section 3.The implementation details are described in Section 4 and finally conclusion is included in the last section.

Related Work:

Many anti-phishing solutions using different technologies have been proposed to predict and tackle phishing attacks. In this section, we review the various solutions and techniques used to identify phishing websites.

An intelligent approach by Mohammad.R.M. *et al*., (2014) is a self-structuring neural network to predict phishing websites.17 features have been used, taking either a binary (phishing or legitimate) or a ternary value (phishing, suspicious, legitimate).The neural network architecture employed in this model is feed forward network. A constructive approach was adopted in specifying the number of neurons in hidden layer. Back propagation algorithm has been used to train the network.

The method employed by He.M *et al*., (2011) determines whether a webpage is either phishing or legitimate, based on its content, HTTP transaction, and search engine results. It is mainly a combination of CANTINA, a content-based approach to detecting phishing web sites; anomalies based web phishing page detection and PILFER a method to detect phishing emails with several additions and modifications. The basic idea of this method is that every website claims a certain identity, and its behavior corresponds to the identity. If a website would be abnormal when compared to a legitimate site, then it can be claimed as fake website. The basis to differentiate between the phishing and legal websites are these abnormalities. This method does not rely on prior knowledge of the user such as web history for phishing detection.

Barraclough.P.A. *et al*., (2013) used a Neuro-Fuzzy Scheme in which 288 features are taken from 5 inputs to detect phishing sites. Two-fold cross validation was used for training and testing. The five inputs that were considered include user behaviour profile (60 features), legitimate site rules (66 features), PhishTank (70 features), user-specific sites(48 features) and pop-up from emails(42 features).

A comparison of machine learning techniques used for detecting malicious web pages is done by Kazemian.H.B. and Ahmed.S., (2015) where two unsupervised machine learning techniques such as k-means and affinity propagation and three supervised machine learning techniques such as the support vector machine, k-nearest neighbor and naive bayes classifier are used. The predictive models have been built using all these machine learning techniques to analyze large number of malicious and safe web pages. Computer simulation results have produced an accuracy of up to 98% for the supervised technique RBF-SVM which is an effective binary classification technique of high dimensional data when compared to other techniques.

A phishing webpage detection approach using Transductive Support Vector Machine (TSVM), a kind of semi supervised learning method is discussed by Li.Y *et al*., (2013) The features of web image are extracted which include gray histogram, color histogram, and spatial relationship between sub graphs. Then the features of sensitive information are checked using page analysis based on Document Object Model. The TSVM trains classifier by considering the distribution information implicitly embodied in the large quantity of the unlabeled samples.

The characteristics of legitimate and phishing web pages were investigated and based on this analysis, heuristic method is proposed by Gowtham.R. and Krishnamurthi.I, (2014) to extract 15 features from web pages. The input to a trained machine learning algorithm SVM to detect phishing sites are these heuristic results. Before applying heuristics to the web pages, two preliminary screening modules are used in this system. The first module, the preapproved site identifier, checks web pages against a private white-list maintained by the user, and the second module, the Login Form Finder, classifies web pages as legitimate when there are no login forms present.

Chen. C.M. *et al*., (2014) presented a feature set that combines the features of social networking and traditional heuristics. Furthermore, a suspicious URL identification system is proposed based on Bayesian classification for use in social network environments. In the first module, data collection, posts are collected including content and time. In the second module, feature extraction, a feature vector is constructed for classification by retrieving the features. In the third module, the Bayesian classification model, a pre trained classification model is used for classifying the posts.

Aghdam.M.H. *et al*., (2009) applied an ACO-based technique to the feature selection problem in text categorization. In this, ACO-based feature selection algorithm has been modified where the performance of the classifier and the length of selected feature subset are taken as heuristic information. A set of experiments was implemented on Reuters-21578 dataset to show the utility of the algorithm and compare it with information gain and CHI (Uysal.A.K. and Gunal.S.,2014).

A novel feature selection algorithm based on rough sets and ACO is proposed by Chen.Y et.al.,(2010) that considers feature significance based on mutual information as heuristic information for ACO. It also introduces the concept of feature core to the algorithm, which requires all ants to start from the core, when they begin their search through the feature space. So, those features near the core will be selected by the ants more quickly. The performance of the algorithm is compared with that of rough set theory based algorithms and other meta heuristic based algorithms. As per the experimental results, RSFSACO has higher performance when compared to other traditional RST-based methods in the ability of finding optimal reduct.

Proposed System:

The overall architecture of the hybrid intelligent system is given in the figure 1.

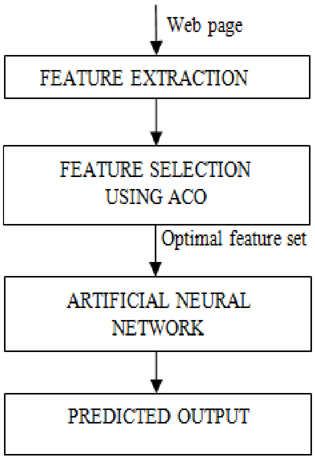

Fig. 1: Overall system architecture

Feature Extraction:

This is the first step in the process and aims to collect features from the webpage. There are several features that distinguish phishing websites from legitimate ones. Features take either a binary or a ternary value. Binary value features hold either ‘‘legitimate’’ or ‘‘phishing’’ as the value is assigned to that feature based on the existence or lack of the feature within the website, whereas for ternary value features, one more value has been added namely ‘‘suspicious’’. For ternary value features, the presence of the feature in a specific ratio determines the value assigned to that feature. Some of these features are listed below:

Table 1: URL Features

S.No	Features
1	IP address
2	Long URL
3	Short URL
4	URL having @ symbol
5	Redirecting using //
6	Prefix and suffix
7	Number of dots in URL
8	URL having user information
9	Sub domain
10	URL of an anchor
11	HTTPS
12	Abnormal URL
13	DNS record lookup
14	Age of domain
15	Longest domain
16	Server form handler

Table 2: Source code features

S.No	Features
1	Pop up window
2	Status bar customization
3	Redirect page
4	Disabling right click
5	Iframe redirection
6	Hiding the links

Aco Based Feature Selection:

The feature selection process using ACO is given in figure 2. The main steps of ACO-based feature selection algorithm are as follows:

- Initialization
- Generation of ants and fitness evaluation
- Selection of subsets
- Checking the stop criteria

Initially, the features are assigned a weight or rank based on the frequency of occurrence of that feature in the phishing url’s dataset [15]. Then, any ant is randomly assigned to one of the features and by visiting other features, each ant builds solutions completely. The subsets of features are gathered along with their fitness evaluation. The criteria for fitness evaluation are based on the contribution of features for predicting the correct output. The subset or reduced feature set is selected such that it the most possible optimal feature set with higher fitness. The stopping criterion is chosen such that either the predefined numbers of iterations are reached or addition or deletion of a feature fails to produce a better subset of features.

Construction Of Artificial Neural Network:

Determining the network architecture is the primary step in the construction of artificial neural network. The neural network architecture used is feed forward with one hidden layer, called the multilayered perceptron. The advantage of this architecture is that the number of neurons in the hidden layer can be changed to adapt to the complication of the relationships between output and input variables.

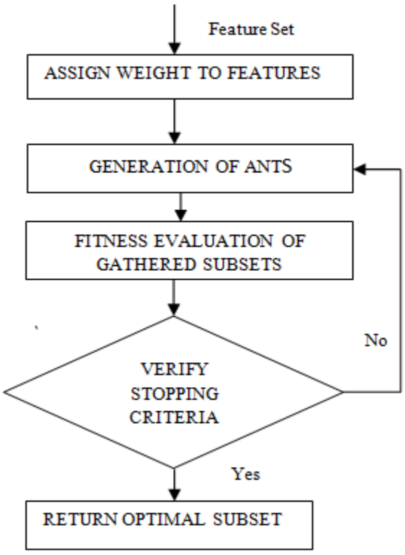

Fig. 2: Ant Colony Optimization based feature selection

In a feed-forward network, data always propagate in one way from input layer to output layer passing through the hidden layers if any. The input layer receives input data from external world, and the neurons in this layer are called input neurons. The input neurons symbolize the data presented to the network for processing. The extracted features are encoded in binary format where value of '1' specifies the presence of the feature and '0' specifies the absence of the feature in url or source code and given as input. The layer following the input layer is the hidden layer, and neurons in this layer are called hidden neurons. The hidden layer receives inputs from the former layer, transforms those inputs into nonlinear combinations and passes the results to the next layer for further processing. The sigmoid activation function is used in all the three layers. Then the training of the network is done where connection weights are adjusted repeatedly until reaching an acceptable solution.

Two learning approaches can be used in the neural networks, namely unsupervised approach and supervised approach. For phishing detection, supervised approach is used since the required output is provided with each training sample. The back propagation learning algorithm is adopted to adjust the network weights. The neural network is constructed in an adaptive manner in accordance with the derived optimal feature set.

Implementation:

The implementation process starts with extracting the features from the webpage. The 16 features in which 12 features are from URL and 4 features from source code are extracted from the web page using a pre-processing code. Then the features are converted to binary format where 1 and 0 specifies the presence and absence of feature respectively by employing the binary encoding scheme so as to provide as input for the artificial neural network. The multilayer perceptron is constructed with 16 input neurons and 2 output neurons in Neuroph studio tool version 2.92. The neural network is illustrated in figure 3.

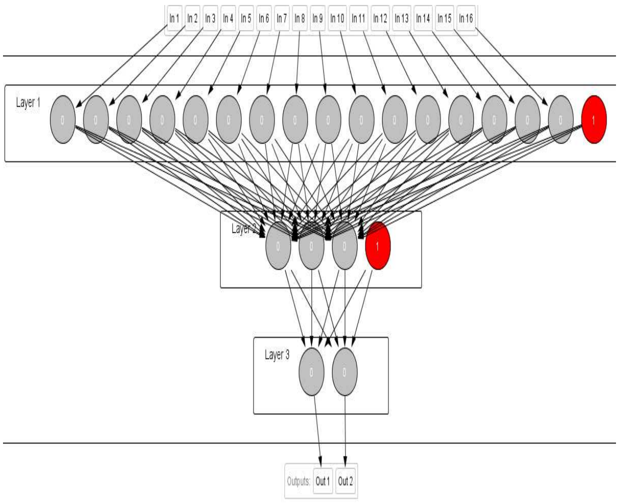

Fig. 3: Structure of the Multilayer Perceptron

The multilayer perceptron is trained with the back propagation learning algorithm. The total network error graph during the learning phase is depicted in figure 4. The predicted output (0,0) specifies legitimate, (0,1) or (1,0) specifies suspicious and (1,1) specifies phishy. The optimal feature set identification using ACO is under implementation.

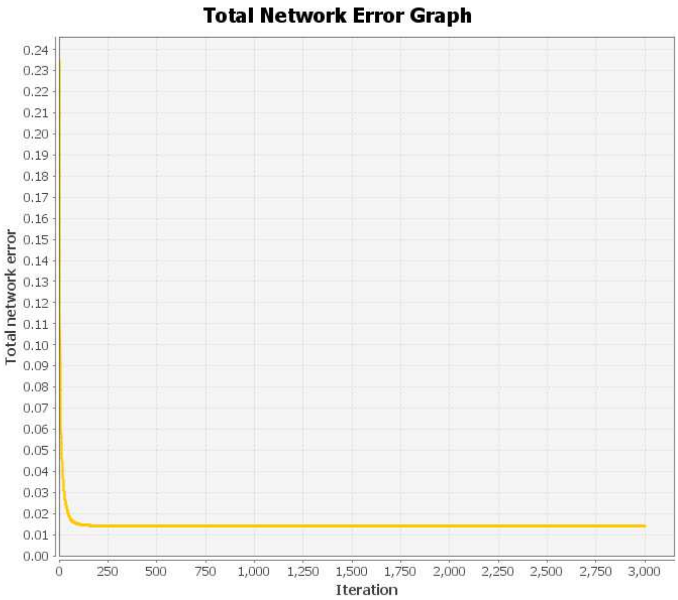

Fig. 4: Total network error graph during the training phase

Conclusion:

A hybrid intelligent approach combining ant colony optimization and artificial neural networks has been proposed to detect phishing web pages. The use of ACO helps to identify optimal or significant features that can detect phishing attacks. This not only helps to reduce the number of features used in the detection process but also help to reduce the complexity of ANN architecture used for further classification. By reducing the complexity in the construction of the ANN, we can improve the time taken for the network to learn and predict the results with higher accuracy.

REFERENCES

Aghdam, M.H., N.G. Aghaee and M.E. Basiri, 2009. Text Feature Selection using Ant Colony Optimization,Expert Systems with Applications, 36(3): 6843-6853.

Barraclough, P.A., M.A. Hossain, M.A. Tahir, G. Sexton and N. Aslam, 2013. Intelligent Phishing Detection and Protection Scheme for Online Transactions, Expert Systems with Applications, 40(11): 4697-4706.

Chen, C.M., D.J. Guan and Q.K. Su, 2014. Feature Set Identification for Detecting Suspicious URLs using Bayesian Classification in Social Networks, Information Sciences, 289: 133-147.

Chen, Y., D. Miao and R. Wang, 2010. A Rough Set Approach to Feature Selection based on Ant ColonyOptimization, Pattern Recognition Letters, 31(3): 226-233.

Das, G., P. Pattnaik and S. Padhy, 2014. Artificial Neural Network trained by Particle Swarm Optimization for non-linear channel equalization, Expert Systems with Applications, 40(7): 3491-3496.

Eglese, R.W., 1990. Simulated annealing: A tool for operational research, European Journal of Operational Research, 46(3): 271-281.

Ganesan, N., K. Venkatesh, M.A. Rama, M.A. Palani, 2010. Application of Neural Networks in Diagnosing Uysal.A.K. and Gunal.S.,2014. The impact of preprocessing on text classification, Information processing and management,15(1):104-112. Cancer Disease Using Demographic Data,International Journal of Computer Application, 26(1): 0975-8887.

Gowtham, R. and I. Krishnamurthi, 2014. A Comprehensive and Efficacious Architecture for Detecting Phishing Web Pages, Computers &amp; Security, 40: 23-37.

Hashimoto, H., T. Kubota, M. Sato, F. Harashima, 1992. Visual control of robotic manipulator based on neural networks, IEEE transactions on Industrial Electronics, 39(6): 490-496.

He, M., S. Horng, P. Fan, M.H. Khan, R.S. Run, J.L. Lai, R.J. Chen and A. Sutanto, 2011. An Efficient Phishing Webpage Detector, Expert Systems with Application, 38(10): 12018-12027.

Kazemian, H.B. and S. Ahmed, 2015. Comparison of Machine Learning Techniques for Detecting Malicious Websites, Expert Systems with Applications, 42(3): 1166-1177.

Li, Y., R. Xiao, J. Feng and L. Zhao, 2013. A Semi-supervised Learning Approach for Detection of Phishing Webpages, Optik - International Journal for Light and Electron Optics, 124(3): 6027-6033.

McCulloch, W.S., W. Pitts, 1943. A logical calculus of the ideas immanent in nervous activity,5(4):115-133.

Mohammad, R.M., F. Thabtah and L. McCluskey, 2014. Predicting Phishing Websites based on Self-Structuring Neural Network, Neural Computing and Applications, 25(2): 443-458.

Uysal, A.K. and S. Gunal, 2014. The Impact of Pre-processing on Text Classification, Information Processing and Management, 15(1): 104-112.

Ramzan, S., 2010. Phishing Attacks and Countermeasures, Handbook of Information and Communication Security, Eds., Stavroulakis.P and Stamp.M,Springer, pp: 433-447.

Socha, K and M. Dorigo, 2008. Ant colony optimization for continuous domains**,** European Journal of Operational Research, 185(3): 1155-1173.