AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

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

Detecting Phishing Urls Using Particle Swarm Optimization

Pradeepthi K.V. and Kannan A.

Department of Information Science and Technology Anna University,Chennai-600 025,Tamil Nadu,India

Address For Correspondence:
Pradeepthi K.V., Department of Information Science and Technology Anna University,Chennai-600 025,Tamil Nadu,India.
E-mail: pradeepthi.kv@gmail.com

ARTICLE INFO

Article history:
Received 04 December 2015
Accepted 22 January 2016
Available online 14 February 2016

Keywords:
Particle swarm optimization; phishing; classifcation; attack detection.

ABSTRACT

The application of evolutionary algorithms to computational problems is very prevalent nowadays because of the way the algorithm is able to adapt to the problem and provide a suitable solution with ease. In this paper, we propose the use of Binary Particle Swarm Optimization (BPSO) technique for detection of phishing URLs. Compared to the existing algorithms used in the literature, we could obtain an improvement in the accuracy and lesser false positive rate. A dataset of 10,000 URLs was constituted and an accuracy of 98.7% was achieved by using this method. By applying, evolutionary techniques to computational problems, we inferred that they solve the problems with better accuracy.

INTRODUCTION

Particle Swarm Optimization (PSO) technique belong to class of evolutionary computing algorithms which try to study the natural scenarios and adapt them to solve various computational problems (Tiago Sousa, 2004). PSO deals with the movement of swarm of birds, typically in search of the nearest source of food. In computational intelligence terminology, this would translate to selecting the best solution from an available set of solutions. Another well known evolutionary computing algorithm is Genetic Algorithm (GA) (Tiago Sousa, 2004). There are many common terminologies in PSO and GA like, population, and fitness value. Population is the set of all possible solutions to a problem. The goal is to select the most valid solution for the problem from the population. Each candidate is called a particle in PSO and it has a fitness value which is calculated by the fitness function. Each particle in PSO has to keep track of two values, pbest(personal best) and gbest(global best). Pbest is the highest value of fitness function achieved by a particular particle throughout, i.e. it is the personal best of particle. Gbest (global best) is the highest fitness value of any particle in the neighborhood. The final goal of the system is to accelerate particle to towards the pbest and gbest values through each iteration. PSO is used for solving computational problems because it can easily adapt to the problem space and move faster towards the solution.

In this paper, we propose the usage of PSO for detection of phishing URLs. The impersonation of a genuine website by a hacker, to trick an innocent end user to click on the given URL is termed phishing. The aim of phishing is to spread malware in to the end user network or perform a fraud by stealing the end user credentials. The rate at which phishing URLs are being generated is alarming and it is not possible for the end users to detect phishing URLs and fake sites without proper detection mechanism in place. A dataset containing a total of 10,000 phishing and fake URLs is built by collecting the URLs from public URL repositories. The different features in the dataset have been selected based on various lexical, network, DNS and URL based features. As all the features in the dataset are binary, we go for binary PSO, which is a variant of the PSO, used when the attributes to be dealt with are binary.

To Cite This Article: Pradeepthi K.V. and Kannan A., Detecting Phishing Urls Using Particle Swarm Optimization. ***Aust. J. Basic & Appl. Sci.,*** *10(2): 75-79, 2016*

In section 2, a detailed literature survey of the different papers in phishing URL detection and about the application of PSO to the web attack detection domain are discussed. In section 3, proposed work and the system architecture are dealt with in detail. Section 4 presents the results and discussions of the proposed system and section talks about the conclusion.

Literature Survey:

Phishing URL detection:

First, the different methodologies used for phishing URL detection are discussed. Zhuang *et al* propose a detection mechanism (Weiwei Zhuang, 2012) where an ensemble of different classification algorithms and finally a clustering algorithm is used to detect phishing URLs. They have analyzed the phishing web site features and extracted the features to generate their own dataset. Hybrid methodologies like using natural language processing techniques are also being used along with the classification algorithms. This can be seen in (Ramanathan, V., H. Wechsler, 2012) that Latent Dirichl*et al*location is used for featuracting features from phishing website and adaboost is used for performing classification. Using heuristics, features in the URL are extracted in (Luong Anh Tuan Nguyen, 2013) and classification is performed. Detection of phishing web sites is done not only by extracting the lexical and heuristic features of the URL but also by employing other techniques like image analysis (Ee Hung Chang, 2013). In this paper, any image of the website is taken and the logo part of the image is segment and compared with the logo of the website present in the Google image database. This comparison yields the result of whether the web site is genuine or phishing. Other than using data mining based techniques, Huang *et al* (2014) have opposed extracting of the features of phishing URLs and using a greedy selection based algorithm for phishing URL detection. Another methodology to detect phishing websites is proposed in (Choon Lin Tan, 2014) is a brand name based weighing system where the domain names present in the website are analyzed and compared with DNS server to detect the authenticity of the website. The phishing URLs are also classified based on their lexical and host based features and their URL ranking (Mohammed Nazim Feroz, Susan Mengel, 2015). The detection of phishing URLs has also been done using the host based, network, URL and lexical features (Pradeepthi, K.V., A. Kannan, 2014).

Application of PSO techniques in attack detection domain:

Since PSO technique is a relatively new concept, the literature in attack detection domain is limited. The competitive nature of PSO in comparison with the existing data mining techniques is discussed in The varied set of problems to which PSO can be applied and the give convening results has been dealt with in (De Falco, I., 2007).

With all these points in mind, we propose a phishing URL detection mechanism using Particle Swarm Optimization with an aim to achieve high detection rate.

Proposed System:

Dataset Generation:

The dataset for the proposed system is collected from public repository DMOZ (…., 2014), which has a large collection of genuine URLs from different domains, manually tested by volunteers. The phishing URLs are collected from the PHISHTANK, which is a collection of phishing URLs. A total of 10,000 URLs are collected, of which 6000 are genuine and 4000 are fake. There are a total of 27 features which belong to various categorized, like lexical, domain based (collected from DNS server), network based and URL feature based.

System Architecture:

The architecture of the proposed system is shown in Figure 1.The generated dataset is first passed to the pre-processing module, where the different features are identified and the missing values are handled.

Then the classification of the URLs is done by using PSO approach for training artificial neural networks. Based on the output got after optimization, the results are sent to the decision module which performs the necessary action for the result received based on the rules detailed out in the rule base. The output of the system is then sent out to the user interface through which the user comes to know about it.

PSO Methodology for Phishing URL Detection:

The PSO algorithm is the new branch of evolutionary algorithm which is being for solving certain computational; problems with ease. It differs from GA because it does not have any operators like cross over and mutation and hence is simpler to implement and execute. It is based on the synergy of a group of birds or school of fish evading a predator of in search of food. Based on the pbest and gbest values of the particles, the full group of particles called swarm travels with a velocity towards the optimal solution.

PSO is used for solving computational problems because it can easily adapt to the problem space and move faster towards the solution.

A fitness function is considered such that

$$R^n \rightarrow R \quad (1)$$

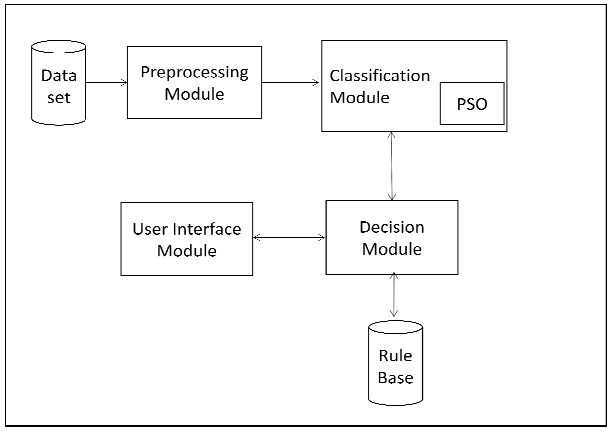

Fig. 1: Proposed system architecture.

This fitness function is used to measure the quality of the solution [13]. Each candidate particle is said to exist randomly in a hyper plane, with position xi, where

$x_i \in R^n$ (2)

The velocity of the particle is given by

$v_i \in R^n$ (3)

The other values that are considered are x_{i+1} which is the changed position of the particle in the next iteration in its attempt to travel towards pbest and gbest, similarly v_{i+1} is the changed velocity of the particle in the next iteration in its attempt to travel towards pbest and gbest, v_{pbest} is the velocity based on pbest and v_{gbest} is the velocity based on pbest. The Figure 2 gives an illustration of how the particle traverses in the hyper plane from the current location to a new location closer to the possible solution based on pbest and gbest values.

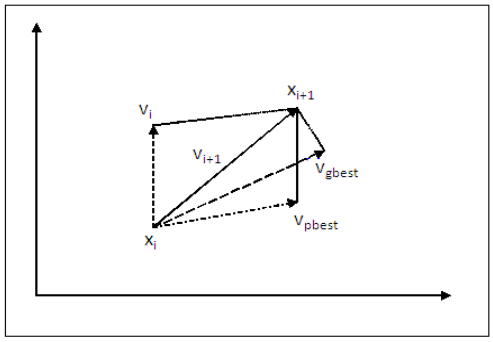

Fig. 2: Searching for the most optimal solution using the PSO algorithm.

The value of the velocity of the particle in the next subsequent iteration to move it towards the best solution is modelled mathematically as given in Equation 4.

$v_i^{k+1} = wv_i^k + c_1 rand_1(\dots)x\left(pbest_i - s_i^k\right) + c_2 rand_2(\dots)x\left(gbest_i - s_i^k\right)$ (4)

Here w is the weight, ci is the weighting factor and $rand_i$ a random number between 0 and 1 which is uniformly distributed.

The value of the weight w is calculated by

$w = w_{initial} - \left[\left(w_{initial} - w_{final}\right)i\right]/\max(i)$ (5)

Where $w_{initial}$ is the initial weight, w_{final} is the final weight and $\max(i)$ is the maximum iteration number possible.

The position of the particle in the next iteration is calculated using

$x_i^{k+1} = s_i^k + v_i^{k+1}$ (6)

Based on the pbest and gbest values of the particles, the full group of particles called swarm travels with a velocity towards the optimal solution. In the proposed methodology, PSO technique is used in training artificial neural networks using matlab.

In our methodology, we used PSO technique for learning in training artificial neural networks using matlab. Each particle is an instance in the dataset and the value of each particle is binary. The number of digits in the binary particle value is equal to the number of features and each digit represents to binary feature value.

Australian Journal of Basic and Applied Sciences, 10(2) Special 2016, Pages: 75-79

The algorithm for PSO is given here.

1: initialize the particles(forming of the binary digits)
2: calculate the fitness function
3: if fitness value > pbest
then
4: assign pbest=fitness value
5: if p(i) = max(pbest)
then
6: max(pbest) = gbest
7: calculate velocity of each particle
8: use velocity value to update particle value
9: if target reached Stop
else
10: Repeat Step 2.

Vital parameter to be determined for this problem is the fitness function for evaluation. It is determined based on the miscalculation rate of the artificial neural network system.

RESULTS AND DISCUSSION

The dataset of 10,000 URLs is split into 5 parts called experiment 1 to experiment 5. The first set has 200 URLs of which 100 are genuine and 100 are fake. The second set has 200 each, the third has 500 each, fourth has 1000 each and the last set has 6000 genuine and 4000 fake URLs.

Table 1: Comparison with Existing Classification Techniques.

Algorithm	Accuracy (%)	False Positive Rate
Naïve Bayes	88.56	0.6
K-Nearest Neighbour	89.4	0.5
ID Tree	92.3	0.45
SVM	91.5	0.58
PSO based System	98.7	0.21

Table 1, compares the accuracy and false positive rate of the different existing classification techniques with the proposed PSO based classification. We can observe that there is an increase in the accuracy of the proposed system. Another important observation is that the false positive rate of the proposed system is low. In the attack detection domain, the FP rate of a system plays an important role because; the malicious URL should never be classified as a genuine one. The FP rate of the different experimental sets is compared in Figure 3 and we can see that as the dataset size increases, the system perform better. The same applies to the precision and recall values shown in Figure2.

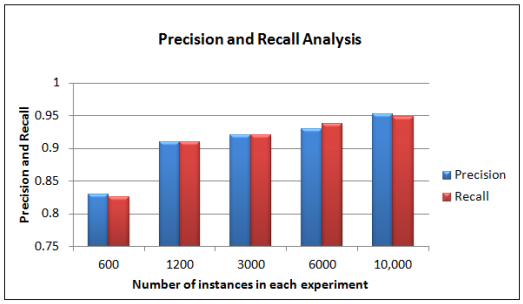

Fig. 3: Graph showing the precision and recall values.

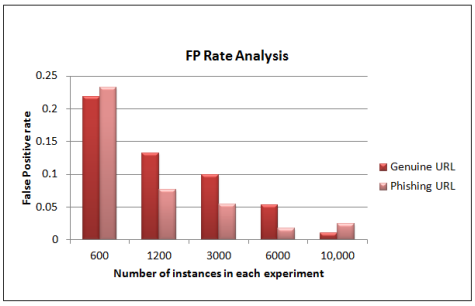

Fig. 4: Graph showing the false positive rate.

Conclusion:

Though the evolutionary algorithms like PSO are being extensively used for solving computational problems, there are not many instances where they are used for solving classification problems. It can be observed from our results that they are able to aide in the training of neural networks and increase the system accuracy while giving lower false positive rate in the compelling field of phishing URL detection. The future enhancement of this paper is to apply PSO along with other classification techniques and analyze the results.

REFERENCES

…, 2014. PhishTank, website. [Online], Available: http://www.phishtank.com.

…., 2014. DMOZ Open Directory Project website. [Online], Available: http://www.dmoz.org.

Choon Lin Tan, Kang Leng Chiew, San Nah Sze, 2014. "Phishing Website Detection Using URL-Assisted Brand Name Weighting System", Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems, 54-59.

Da Huang, Kai Xu, Jian Pei, 2014. "Malicious URL detection by dynamically mining patterns without pre-defined elements", World Wide Web, 17(6): 1375-1394.

De Falco, I., A. Della Cioppa, E. Tarantino, 2007. "Facing classification problems with Particle Swarm Optimization", Applied Soft Computing, 7: 658-665.

Ee Hung Chang, Kang Leng Chiew, San Nah Sze, Wei King Tiong, 2013. "Phishing Detection via Identification of Website Identity", Proceeing of Internaional Conference of IT Convergence and Security, 1-4.

Kolias, C., G. Kambourakis, M. Maragoudakis, 2011. " Swarm intelligence in intrusion detection: A survey", Computers and Security, 30(8): 625-642.

Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen, Minh Hoang Nguyen, 2013. "Detecting phishing web sites: A heuristic URL-based approach", Proceedings of International Conference on Advanced Technologies for Communications, 597-602.

Mohammed Nazim Feroz, Susan Mengel, "Phishing URL detection using URL Ranking", Proceeidng of International Congress on Big Data, pp. 635-638, 2015.

Pradeepthi, K.V., A. Kannan, 2014. “Performance Study of Classification Techniques for Phishing URL Detection”, 6th IEEE International Conference on Advanced Computing, 135-139.

Ramanathan, V., H. Wechsler, 2012. "Phishing Website Detection Using Latent Dirichl*et al*location and Adaboost", Proceedings of International Conference on Intelligence and Security Informatics, 102-107.

Tiago Sousa, Arlindo Silva, Ana Neves, 2004. "Particle Swarm based Data Mining" Parallel Computing, 30 (5): 767-783.

Weiwei Zhuang, Qingshan Jiang, Tengke Xiong, 2012. "An Intelligent Anti-phishing Strategy Model for Phishing Website Detection", Proceedings of International Conference on Distributed Computing System Workshops, 51-56.