IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Phishing URL Detection using Neural Network Optimized by Cultural Algorithm

A. Haider[1*], R. Singh[2]

[1]Dept. Computer Science, IEC Group of Institutions, AKTU, Greater Noida, India
[2]Dept. Computer Science, IEC Group of Institutions, AKTU, Greater Noida, India

[*]*Corresponding Author: atebar.haider786@gmail.com, Tel.: +91 9918640979*

Available online at: www.ijcseonline.org

Accepted: 10/Jul/2018, Published: 31/Jul/2018

***Abstract*—** Internet scams are numerous and varied. Anyone is likely to be the target of an attack while browsing the net. More and more crooks do not hesitate to use Social Engineering as a lever to acquire sensitive data unfairly by exploiting human flaws. Phishing is a Social Engineering technique used by these hackers. It is used to steal personal information in order to commit an identity theft without the knowledge of their victims. The persuasion power of these crooks is the keystone of a successful attack. This work aims to collect, map and model elements that will lead to the finding of phishing URL automatically, for this purpose data mining is used as basic tools, in this sense, it is considered that the existing patterns in a URL make it possible to distinguish the legitimate link for pages, the identification of these patterns will serve to model a successful classification method, for this purpose, the attributes found in the database "phishing web" that correspond to patterns of phishing pages will be validated, at the same time will be evaluated algorithms extracted from the literature that allow a better classification of records, finally, a model with the highest precision results is delivered which consists of cultural algorithm optimized neural network classifier.

***Keywords*—** Cultural Algorithm, Neural Network, Phishing URL

I. INTRODUCTION

Phishing is an attempt to steal personal confidential information such as passwords, credit card information from innocent victims for financial gain, identity theft and other fraudulent activities by an individual or a group. The current scenario, when the user desires to access his confidential information online (like payment gateway or money transfer) by logging into his secure mail account or bank account, the individual enters information like credit card no., username, password etc. on the login page. But quite often, this information can be taken by intruders using phishing techniques (for example, when a user provides login information on a phishing website his data is stolen and then he is redirected to the genuine site). There is no such information that cannot be directly obtained from the user at the time of his login input.

Whittaker et al. [1] define a phishing web page as "any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewers would only trust a true agent of a the third party."

Phishing is a generally a web criminality in relationship with various structures, for example, virus attacks and hacking. In recent times, an expansive number of phishing web pages have been discovered. Its effect is data security rupture through the cooperation of classified information and the objectives may at long last endure loss of cash or different types.

Phishing web pages are fake web pages that are made by malicious individuals to mimic Web pages of genuine web sites. Most of these types of web pages have great visual similarities to trick their victims. Some of these types of web pages look exactly like the genuine ones. Victims of phishing web pages may expose their credit card number, password, bank account or other vital information to the phishing web page owners. It includes techniques such as deceiving customers through URL, screen captures, spam messages, emails and installation of key loggers.

This paper focuses on a lexical analysis of URLs because they are:

- Less expensive to process than external information or content-data.
- URLs are more likely to be stored and obtainable as they use up fewer resources, such as disk space than external information and content-data. Content-based analysis more costly to obtain.
- There is no point to using complex operations if we have not evaluated simpler operations. If simpler operations are not productive enough, then we may want to use more complex operations. However, as covered in subsequent sections, our study shows that a lot can be achieved with just lexically analysing URL.

The main objective of this paper is:

- To develop a framework for classification and detection of phishing URL using Neural Network and Cultural Algorithm based approaches.

Performance of the proposed research work will be carried out using certain evaluation parameters, namely; Accuracy, Recall, Precision, False Negative Rate, False Positive Rate, True Positive Rate and True Negative Rate.

II. ANTI-PHISHING TECHNOLOGY

Tools for anti-phishing offer users an active scheme which warns and protect against likely phishing scams, also they help us guard the brand image of genuine ISPs as well as the developers of e-commerce site to be "spoofed" to spread scams. Certainly, the best significant part of tools which detect phishing is they recognize malicious websites in a significantly precise manner, also in an adequate time span. Nearly most of these tools offer binary indicators showing if a site is genuine or not, which could be accomplished by making use of colourful indications (green symbolises a genuine site, in addition to this red symbolises a positively identified phishing website). Further these tools practice a system which characterizes the website, into phishing, genuine or indefinite (doubtful), also this could be executed by making use of colourful indications (green signifies genuine site, red signifies positively identified phishing website lastly, and a grey or yellow pointer signifies an indefinite website which is suspicious to phishing attack).

Phishing techniques, these days, have not only increased in quantity, but also in complexity [2]. To carry out an ingenious phishing attack phishers make use of lots of tactics and methods. Getting on to users of Internet-based services like, on-line banking users and providers of payment services are the major targets of the phishing scams are dealing with a huge amount of loss of trust and financial loss. There is a serious want to uncover way out to overcome the phishing scams. Until now, numerous fixes are being offered and implemented in course of phishing. These fixes focus on both technical problem and non-technical scenarios.

Several ideas were borrowed from Spoofguard and additional checks were added to figure out the trends within the phishing websites. However, in spite of different scenarios it is difficult to provide maximum accuracy. The core problem is to decrease the detection of false positives and increase the true positives thereby increase the overall accuracy of the system.

The major concern of this research is to design a framework intended for assessment of the lexical features to show signs of improvement through comprehensively studying the components of the URLs which promote phishing, by the means of Neural Network and Cultural Algorithm based classifying algorithms.

III. METHODOLOGY

A. Proposed Architecture

The classifier takes unclassified URLs as input, and returns a predicted binary class as output (either Phish or Benign). Our aim is to evaluate the effectiveness of URL features as discriminating features.

We started with collection of URLs and then after loading the URLs we started by reading URLs one by one for feature extraction.

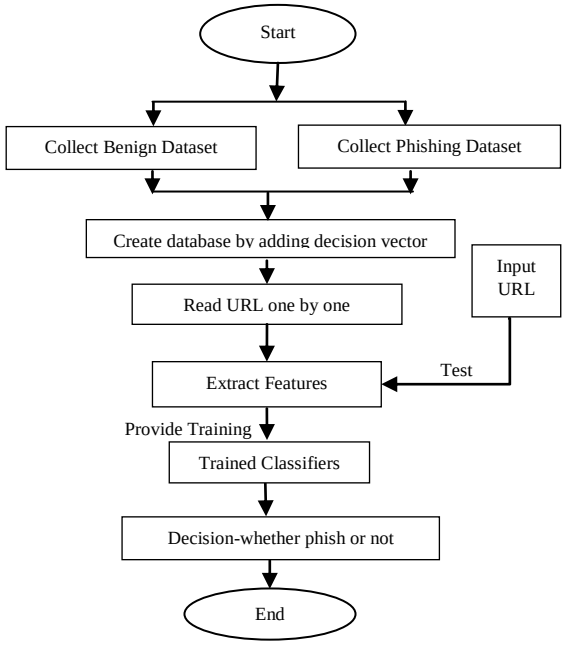

Figure 1: Flow diagram of proposed architecture

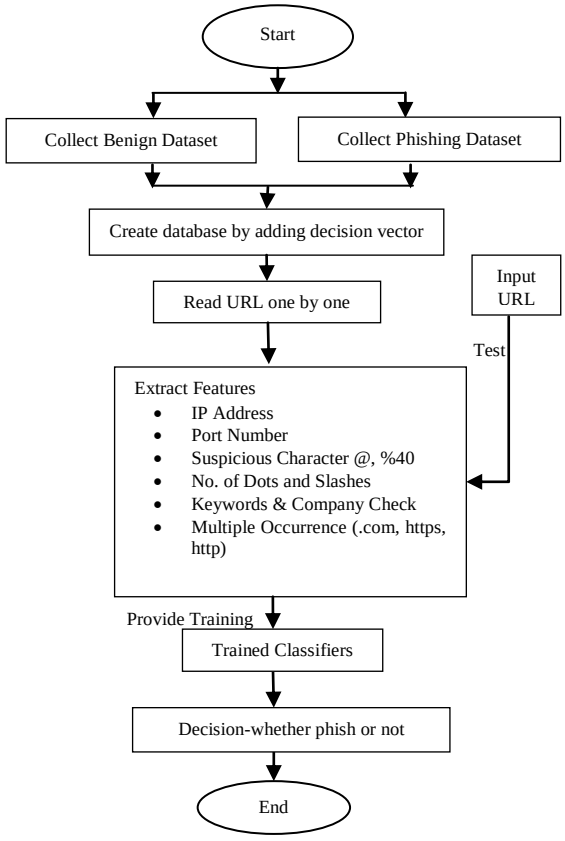

Figure 2: Flow diagram for lexical feature extraction

To facilitate feature extraction, each URL was split into three sections: protocol, domain, and path. All subsequent feature extraction was performed on these sub-regions. After collecting of URL features, the classifier's life initiates by a supervised learning phase. During this phase, the classifier is fed with pre-classified URL along with their pre-defined class. The classifier is then able to perceive a classification model. Once the learning phase is complete, the classifier is given unclassified URLs as input, and a predicted class is returned as output.

B. Collection of URLs

Here in this research work, we have taken URLs of benign websites from www.alexa.com, www.dmoz.org, and personal web browser history. The phishing URLs were collected from www.phishtak.com.

C. Lexical Feature Extraction

Lexical features are the textual properties of the URL itself, not the substance of the page it indicates [3]. URLs are human-readable text strings that are parsed in a standard manner by customer projects. Through a multistep determination process, programs make an interpretation of each URL into guidelines that find the server facilitating the site and indicate where the site or asset is set on that host.

- IP Address
- Protocol
- Number of Dots and Slashes
- Suspicious Character @ and %40
- Multiple Occurrence (.com, https, http)
- Keyword Check
- Company Check

D. Classification Algorithms

The input to the classifiers in MATLAB is two .txt files; newben.txt and newphis.txt. The two classification algorithms considered for processing the feature set are:

1) Neural Network (NN)

The Artificial Neural Network (ANN) is the replica of animal's central nervous system specifically designed to meet the interests of machine learning for pattern recognition. Back Propagation Neural Network (BPNN) generates complex decision boundaries in feature space. BPNN in specific circumstances resembles Bayesian Posterior Probabilities at its output. These conditions are essential to achieve low error performance for given set of features along with selection of parameters such as training samples, hidden layer nodes and learning rate. In else case, the performance of BPNN could not be evaluated. For W number of weights and N number of nodes, numbers of samples (m) are depicted to correctly classify future samples in following manner [4]:

$$m \geq O\left(\frac{W}{\epsilon}\log\frac{N}{\epsilon}\right) \quad (1)$$

The theoretical computation of number of hidden nodes is not a specific process for hidden layers. Testing method is commonly entertained for selection of these followed in the constrained environment of performance.

2) Cultural Algorithm (CA)

Inspired by the process of social and cultural changes, the CA was developed to enhance evolutionary computation. Besides the population component that evolutionary computation approaches have, there is an additional peer component belief space and a supporting communication protocol between these two components, which makes CAs perform better in some special optimal cases than other evolutionary algorithms (EAs). The following figure presents the basic CA framework.

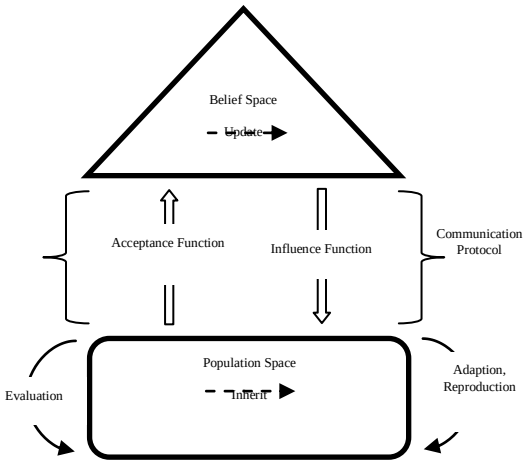

Figure 3: CA framework [6]

As Figure 3 shows, the population space and the belief space can evolve respectively. The population space consists of the autonomous solution agents and the belief space is considered as a global knowledge repository. The evolutionary knowledge that stored in belief space can affect the agents in population space through influence function and the knowledge extracted from population space can be passed to belief space by the acceptance function.

The CA pseudo code presented by [5] is given as follows:

```
t=0;
Initialize Population POP(t);
Initialize Belief Space BLF(t);
Repeat
    Evaluate Population POP(t);
    Adjust (BLF(t), Accept(POP(t)));
 Adjust (BLF (t));
 Variation(POP (t) from POP (t-1));
Until termination condition achieved
```


IV. RESULTS AND DISCUSSION

The performance of proposed algorithms has been studied by means of MATLAB simulation.

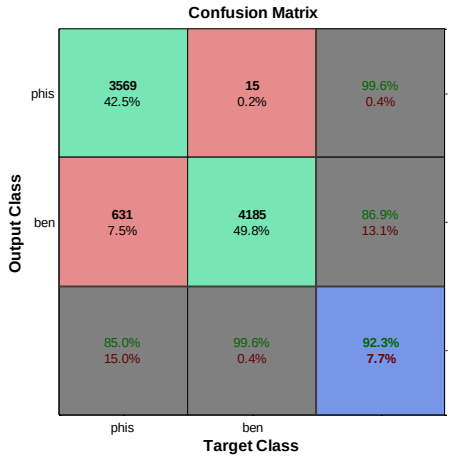

Figure 4: Confusion matrix for Neural Network classifier

The confusion matrix plot indicates accuracy i.e. 92.3% for this approach.

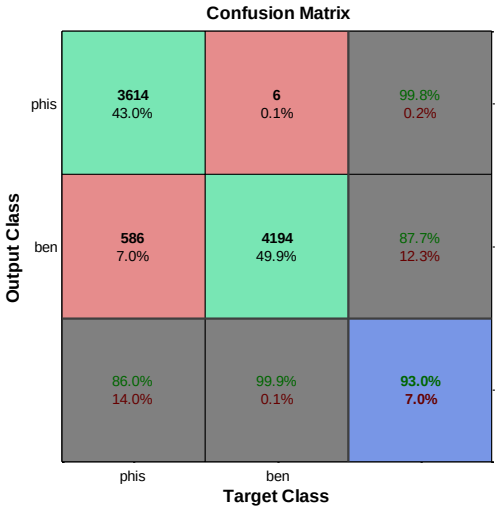

Figure 5: Confusion matrix for cultural optimized neural network algorithm

The confusion matrix plot indicates accuracy i.e. 93.0% for this approach.

V. CONCLUSION

Phishing recognition techniques are rapidly varying to keep up with the novel techniques used by phishers. Combating phishing is an on-going battle that will perhaps never end much like the on-going battle with spam emails. Phishers have various methodologies and procedures to conduct a well-designed phishing attack.

While generalizing about URLs, it is hard to conclude if a website is genuine or phishing just by the contents of the URL alone. One can on the other hand add to a phishing score if certain features are spotted that are more likely found in phishing URLs rather than legitimate URLs.

We have made use of simple techniques for classification as our intention was the evaluation of the feature, and not the classifiers. This work proved diagnostically that the proposed methodology is showing the signs of improvement utilizing different lexical features for detecting phishing URLs through proposed classifier.

REFERENCES

[1] C. Whittaker, B. Ryner and M. Nazif, "*Large-scale automatic classification of phishing pages*," in In: Proc. 17th Annual Network and Distributed System Security Symposium, NDSS"10, San Diego, CA, USA, 2010.

[2] Meenu Shukla, Sanjiv Sharma, "*Analysis of Efficient Classification Algorithm for Detection of Phishing Site*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.3, pp.136-141, 2017

[3] Xiang G., Hong J., Rose C. P. and Cranor L. , "*CANTINA+: A feature-rich machine learning framework for detecting phishing Web sites*," ACM Trans. Inf. Syst. Secur. 14, 2, Article 21, p. 28, September 2011.

[4] Christos Stergiou and Dimitrios Siganos, "*Neural Networks*", Report available at: http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[5] Reynolds, Robert G. "*An introduction to cultural algorith*ms." In Proceedings of the third annual conference on evolutionary programming, pp. 131-139. Singapore, 1994.

