IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 09, April 2019)

Phishing URL Detection using PSO Optimized Support Vector Machine

Ankit Shrivastava
M. Tech. Scholar
Dept. of Electronics and Communication
Sagar Institute of Research and Technology, Indore,
M.P. (India)

Dr. Sudhir Agrawal
Dean academics
Sage University Indore, M.P. (India)

Abstract – **This paper 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 particle swarm optimized support vector machine classifier.**

Keywords – **Benign, Phish, Phishing URL, PSO, SVM.**

I. INTRODUCTION

Phishing fraud - that is, the theft of banking or personal information by phishing techniques and their conversion into money or goods and services - has been steadily increasing for several years and the phenomenon does not seem to have occurred. On the contrary, it has become a widespread practice among web crooks due to the increased use of social networks, e-commerce, mobile devices [1], [2] and cloud solutions to store and manage sensitive data [3]. To convince oneself of this, simply type the terms "bank", "fraud", "Scam" and "phishing" in Google or Google Scholar. We get almost a million results in Google and 10,800 in Google Scholar. This is how important the topic is on the Internet and arouses the interest of researchers and organizations that fight against phishing. Among these organizations, there is the Anti-Phishing Working Group (APWG) which published in its 2017 report that more than 91% of all phishing attacks in 2016 targeted five types of industries in particular, financial institutions, cloud-based data hosts, web hosts, online payment services and e-commerce services. This figure of 91% represents an average increase of 33% per type of industry compared to 2015. An increase which is, however, abnormally high for Canadian companies that have experienced, among the developed countries, the strongest phishing growth in 2016, nearly 237% according to the Phishlabs 2017 report [4], mainly in the financial institutions sector, where the 444% ceiling was reached [5-8]. Trademarks targeted by phishing campaigns reached an average 2016 record of 380 per month, 13% higher than the previous year.

In addition to targeting businesses and trademarks, fraudsters target consumers who connect to the Internet [9-11].

In summary, what we can learn from these numbers is that phishing fraud:

- Makes more and more victims;
- In the short term, at the individual level, losses and undermines confidence in the Internet for online transactions. And, at the corporate level, it would undermine the trust of customers and account holders in them and damage their images [12].
- Adapts more and more to new information technologies (e.g. SMS);
- Target new economic sources such as companies that manage information.

The real worry of this exploration is to outline a system expected for appraisal of the lexical features to show signs of improvement through comprehensively studying the components of the URLs which promote phishing, by the methods of particle swarm optimized support vector machine classifier.

The main objective of this paper is to develop a framework for classification and detection of phishing URL using SVM and PSO optimized SVM classifier approaches. Performance of the proposed research work is carried out using certain evaluation parameters, namely; Accuracy, Recall, Precision,

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 09, April 2019)

False Negative Rate, False Positive Rate, True Positive Rate and True Negative Rate

II. Proposed 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.

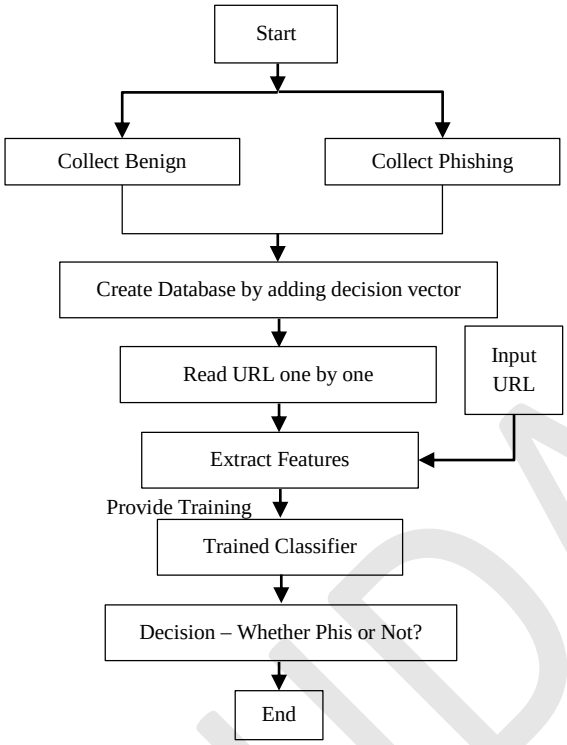

Figure 1: Flow diagram of proposed architecture

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

Architectures also hold room for checking a particular URL for Phishing. A random URL is provided to the trained classifier for recognizing the class (Phishing or Benign) of the given URL.

B. Collection of URLs

Here in this research work, we have taken URLs of benign websites from www.alexa.com [13] www.dmoz.org [14] and personal web browser history. The phishing URLs were collected from www.phishtak.com [15].

C. Lexical Feature Extraction

Lexical features are the textual properties of the URL itself, not the substance of the page it indicates. 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 classification algorithm considered for processing the feature set is:

Particle Swarm Optimization of Support Vector Machine

A direction for PSO is to optimize continuous and mixed (discrete and continuous) variables in solving problems with various types of data. Support Vector Machine (SVM), which originates from the statistical approach, is a present day classification technique. The main problems of SVM are selecting feature subset and tuning the parameters. Discretizing the continuous value of the parameters is the most common approach in tuning SVM parameters. This process will result in loss of information which affects the classification accuracy. This proposed work discuss the algorithm that can tune SVM parameters.

The algorithms that are proposed in this work are related to optimizing two SVM parameters. The parameters are: i) weight, C; and ii) kernel function. The weight represents the trade-off between misclassifying certain points and correctly classifying others, while the kernel is used to simultaneously tune SVM parameters and select the feature subset.

PSO-SVM Algorithm

Input: k, m, q, C, γ, and termination criterion

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 09, April 2019)

Output: Optimal value for SVM parameters and classification accuracy
Begin
Initialize k solutions
call SVM algorithm to evaluate k solutions
T = Sort (S1, …, Sk)
while classification accuracy ≠ 100% or number of iteration ≠ 10 do
for i = 1 to m do
select S according to its weight
sample selected S
store newly generated solutions
call SVM algorithm to evaluate newly generated solutions
end
T = Best (Sort S1, … Sk + m), k)
end
end

In the algorithm, k is the size of solution archive, m is the number of swarm that are used to generate solutions, q is the algorithm's parameter to control diversification of the search process, C is the regularization or soft margin parameter, γ is the kernel function parameter called the margin or the width parameter, and finally, the termination conditions for the best values for SVM parameters (C and γ).

III. Simulation Results

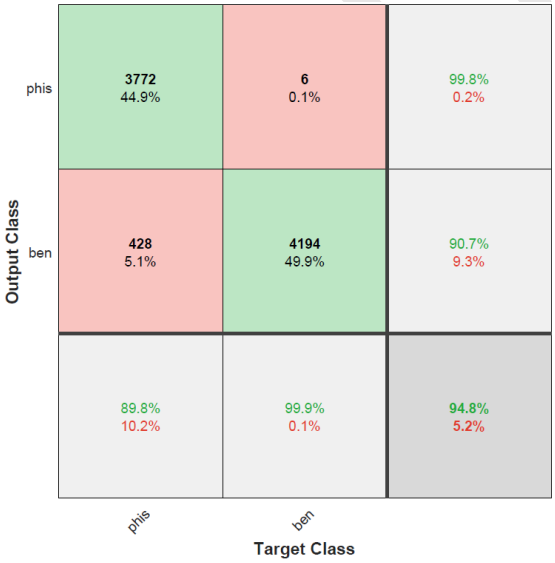

Figure 2: Confusion matrix plot for SVM classifier based method

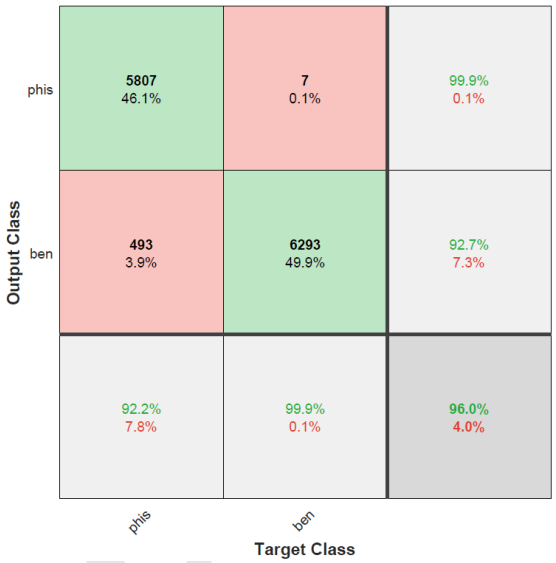

Figure 3: Confusion matrix plot for particle swarm optimized support vector machine classifier method

IV. Conclusion

The results presented in the presented model require obtaining results with a higher level of accuracy, since the need in terms of safety should be close to 100% with a fault tolerance of 0.001%.
The database "phishing web" offers a number and variety of attributes established by all the literature, however, the tests carried out show that of the 60-40 split case is presented in the simulation, nevertheless, it is proposed a possible consensus on the attributes that can come to clearly define a phishing URL. On the other hand, the amount of consigned attributes turns out to be an inconvenience due to the "curse of the dimension", since understanding and processing all these attributes translates into space, time and costs.
In this paper, the gain that occurs when using classification techniques such as particle swarm optimized support vector machine classifier is revealed at the theoretical level, even though no technique is superior to the others in a general way, since they have limitations and own advantages that are coupled according to the model we are working with.

Reference

[1] Jaeger, J.-M. D. (2016). Des pirates volent 72 millions de dollars à une plateforme de Bitcoin. Retrieved from http://www.lefigaro.fr/secteur/high-tech/2016/08/03/32001-20160803ARTFIG00143-des-pirates-volent-72-millions-de-dollars-a-une-plateforme-de- bitcoin.php

[2] Lastdrager, E. E. (2014). Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Science*, *3*(1), 9.

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 09, April 2019)

[3] Legaldictionary. (2017). Bank Fraud. Retrieved from https://legaldictionary.net/bank-fraud/
[4] Phishlabs. (2017). *2017 Phishing Trends and Intelligence Report: Hacking the Human*. Retrieved from https://pages.phishlabs.com/rs/130-BFB-942/images/2017%20PhishLabs%20Phishing%20and%20Threat%20Intelligence%20Report.pdf
[5] Ramsey, D. (2017). Bank Fraud Law and Legal Definition. Retrieved from https://definitions.uslegal.com/b/bank-fraud/
[6] Chan, T. (2004). HK$660,000 stolen in e-bank scam. *China Daily HK Edition.* Retrieved from http://www.chinadaily.com.cn/english/doc/2004-10/08/content_380368.htm
[7] CHAWKI, M. (2006). Phishing in Cyberspace: Issues and Solutions. Retrieved from http://www.crime-research.org/articles/phishing-in-cyberspace-issues-and-solutions/
[8] Renaudin, K. (2011). *Le spamming et le droit: analyse critique et prospective de la protection juridique des" spammés"*. Université de Grenoble,
[9] Mihai, I.-C. (2012). Overview on Phishing Attacks. *Int'l J. Info. Sec. & Cybercrime*, *1*, 61. Milletary, J., & Center, C. C. (2005). Technical trends in phishing attacks. *Retrieved December*, *1*(2007), 3.3.
[10] Hutchings, A., & Holt, T. J. (2017). The online stolen data market: disruption and intervention approaches. *Global Crime*, *18*(1), 11-30.
[11] Chhikara, J., Dahiya, R., Garg, N., & Rani, M. (2013). Phishing & anti-phishing techniques: Case study. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(5).
[12] Symantec. (2016). *Internet Security Threat Report*. Retrieved from https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf
[13] "The Web Information Company," [Online]. Available: www.alexa.com.
[14] "DMOZ Open Directory Project," [Online]. Available: http://www.dmoz.org.
[15] "PhishTank," [Online]. Available: https://www.phishtank.com/