Phishing web sites features classification based on extreme learning machine

Conference Paper · March 2018

DOI: 10.1109/ISDFS.2018.8355342

Yasin Sonmez
Batman University

Turker Tuncer
Fırat University

Hüseyin Gökal
İstanbul Esenyurt Universty

Phishing Web Sites Features Classification Based on Extreme Learning Machine

Yasin Sönmez[1]
Dicle University -Technical Sciences Vocational School
Diyarbakır / Turkey
yasin.sonmez@dicle.edu.tr

Türker Tuncer[2]
Fırat University-Faculty of Technology Forensic Comp.
Elazığ / Turkey
turkertuncer@firat.edu.tr

Hüseyin Gökal [3]
Cyprus International University Faculty of Edu.
Lefkoşa / Cyprus
hgokal@ciu.edu.tr

Engin Avcı[4]
Fırat University-Faculty of Technology Sofware Eng.
Elazığ / Turkey
enginavci@firat.edu.tr

***Abstract*—Phishing are one of the most common and most dangerous attacks among cybercrimes. The aim of these attacks is to steal the information used by individuals and organizations to conduct transactions. Phishing websites contain various hints among their contents and web browser-based information. The purpose of this study is to perform Extreme Learning Machine (ELM) based classification for 30 features including Phishing Websites Data in UC Irvine Machine Learning Repository database. For results assessment, ELM was compared with other machine learning methods such as Support Vector Machine (SVM), Naïve Bayes (NB) and detected to have the highest accuracy of 95.34%**

Keywords—Extreme Learning Machine,Features Classification, Information Security, Phishing.

I. Introduction

Internet use has become an essential part of our daily activities as a result of rapidly growing technology. Due to this rapid growth of technology and intensive use of digital systems, data security of these systems has gained great importance. The primary objective of maintaining security in information technologies is to ensure that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies [1]. Phishing is defined as imitating reliable websites in order to obtain the proprietary information entered into websites every day for various purposes, such as usernames, passwords and citizenship numbers. Phishing websites contain various hints among their contents and web browser-based information [2-4]. Individual(s) committing the fraud sends the fake website or e-mail information to the target address as if it comes from an organization, bank or any other reliable source that performs reliable transactions. Contents of the website or the e-mail include requests aiming to lure the individuals to enter or update their personal information or to change their passwords as well as links to websites that look like exact copies of the websites of the organizations concerned [6-10].

Phishing Web sites Features

Many articles have been published about how to predict the phishing websites by using artificial intelligence techniques. We examined phishing websites and extracted features of these web sites. Guidelines regarding the extracted features of this database are given below.

In the first section we defined rules and we gave equations of web features. We need these equations in order to explain phishing attacks characaterization.

1.1. Address Bar based Features

1.1.1. Using the IP Address

<u>Rule</u>:

$$\begin{cases} \text{If The Domain Part has an IP Address} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{1}$$

1.1.2. Long URL to Hide the Suspicious Part

$$\begin{cases} URL\ length < 54 \rightarrow feature = \text{Legitimate} \\ else\ if\ URL\ length \geq 54\ and\ \leq 75 \rightarrow feature = Suspicious \\ otherwise \rightarrow feature = \text{Phishing} \end{cases} \tag{2}$$

1.3. Using URL Shortening Services “TinyURL”

$$\begin{cases} \text{TinyURL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{3}$$

1.1.4. URL’s having “@” Symbol

$$\begin{cases} \text{TinyURL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{4}$$

1.1.5. Redirecting using “//”

$$\begin{cases} \text{ThePosition of the Last Occurrence of "//" in the URL} > 7 \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{5}$$

1.1.6. Adding Prefix or Suffix Separated by (-) to the Domain

$$\begin{cases} \text{Domain Name Part Includes } (-) \text{ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \tag{6}$$

1.1.7. Sub Domain and Multi Sub Domains

$$\begin{cases} \text{Dots In Domain Part} = 1 \rightarrow \text{Legitimate} \\ \text{Dots In Domain Part} = 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (7)$$

1.1.8.HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

$$\begin{cases} \text{Use https and Issuer Is Trusted and Age of Certificate} \geq 1 \text{ Years} \rightarrow \text{Legitimate} \\ \text{Using https and Issuer Is Not Trusted} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (8)$$

1.1.9. Domain Registration Length

$$\begin{cases} \text{Domains Expires on} \leq 1 \text{ years} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (9)$$

1.1.10. Favicon

$$\begin{cases} \text{Favicon Loaded From External Domain} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (10)$$

1.1.11. Using Non-Standard Port

$$\begin{cases} \text{Port \# is of the Preffered Status} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (11)$$

1.1.12. The Existence of "HTTPS" Token in the Domain Part of the URL

$$\begin{cases} \text{Using HTTP Token in Domain Part of The URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (12)$$

1.2. Abnormal Based Features

1.2.1. Request URL

$$\begin{cases} \text{\% of Request URL} < 22\% \rightarrow \text{Legitimate} \\ \text{\%of Request URL} \geq 22\% \text{ and } 61\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{feature} = \text{Phishing} \end{cases} \quad (13)$$

1.2.2 URL of Anchor

$$\begin{cases} \text{\% of URL Of Anchor} < 31\% \rightarrow \textit{Legitimate} \\ \text{\% of URL Of Anchor} \geq 31\% \text{ And } \leq 67\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (14)$$

1.2.3 Links in <Meta>, <Script> and <Link> tags

$$\begin{cases} \text{\% of Links in "} <\text{Meta}> \text{","} <\text{Script}> \text{" and "} <\text{Link}> \text{"} < 17\% \rightarrow \text{Legitimate} \\ \text{\% of Links in } <\text{Meta}> \text{","} <\text{Script}> \text{" and "} <\text{Link}> \text{"} \geq 17\% \text{ And} \leq 81\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (15)$$

1.2.4. Server Form Handler (SFH)

$$\begin{cases} \text{SFH is "about: blank" Or Is Empty} \rightarrow \text{Phishing} \\ \text{SFH Refers To A Different Domain} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (16)$$

1.2.5. Submitting Information to Email

$$\begin{cases} \text{Using "mail()" or "mailto:" Function to Submit User Information} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (17)$$

1.2.6.Abnormal URL

$$\begin{cases} \text{The Host Name Is Not Included In URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (18)$$

1.3 HTML and JavaScript based Features

1.3.1. Website Forwarding

$$\begin{cases} \text{\#ofRedirect Page} \leq 1 \rightarrow \text{Legitimate} \\ \text{\#of Redirect Page} \geq 2 \text{ And } < 4 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (19)$$

1.3.2 Status Bar Customization

$$\begin{cases} \text{onMouseOver Changes Status Bar} \rightarrow \text{Phishing} \\ \text{It Does't Change Status Bar} \rightarrow \text{Legitimate} \end{cases} \quad (20)$$

1.3.3. Disabling Right Click

$$\begin{cases} \text{Right Click Disabled} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (21)$$

1.3.4. Using Pop-up Window

$$\begin{cases} \text{Popoup Window Contains Text Fields} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (22)$$

1.3.5. IFrame Redirection

$$\begin{cases} \text{Using iframe} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (23)$$

1.4. Domain based Features

1.4.1. Age of Domain

$$\begin{cases} \text{Age Of Domain} \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (24)$$

1.4.2.DNS Record

$$\text{Rule: IF} \begin{cases} \text{no DNS Record For The Domain} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (25)$$

1.4.3.Website Traffic

$$\begin{cases} \text{Website Rank} < 100{,}000 \rightarrow \text{Legitimate} \\ \text{Website Rank} > 100{,}000 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (26)$$

1.4.4. PageRank

$$\begin{cases} \text{PageRank} < 0.2 \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (27)$$

1.4.5. Google Index

$$\begin{cases} \text{Webpage Indexed by Google} \rightarrow \text{Legitimate} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases} \quad (28)$$

1.4.6. Number of Links Pointing to Page

$$\begin{cases} \text{\#Of Link Pointing to The Webpage} = 0 \rightarrow \text{Phishing} \\ \text{\#Of Link Pointing to The Webpage} > 0 \text{ and} \leq 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (29)$$

1.4.7.Statistical-Reports Based Feature

$$\begin{cases} \text{Host Belongs to Top Phishing IPs or Top Phishing Domains} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases} \quad (30)$$

In this study, Extreme Learning Machine (ELM) based classification was performed for the following 30 features [11] extracted based on the features of websites in UC Irvine

Machine Learning Repository. In the Table 1, features of web sites are listed.

TABLE I. FEATURES OF WEBSITES

Input (Features)	*Output (Class)*
1.1. Address Bar based Features 1.1.1. Using the IP Address 1.1.2. Long URL to Hide the Suspicious Part 1.1.3. Using URL Shortening Services “TinyURL” 1.1.4. URL’s having “@” Symbol 1.1.5. Redirecting using “//” 1.1.6. Adding Prefix or Suffix Separated by (-) to the Domain 1.1.7. Sub Domain and Multi Sub Domains 1.1.8. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer) 1.1.9. Domain Registration Length 1.1.10. Favicon 1.1.11. Using Non-Standard Port 1.1.12. The Existence of “HTTPS” Token in the Domain Part of the URL	
1.2. Abnormal Based Features 1.2.1.Request URL 1.2.2.URL of Anchor 1.2.3.Links in <Meta>, <Script> and <Link> tags 1.2.4.Server Form Handler (SFH) 1.2.5.Submitting Information to Email 1.2.6.Abnormal URL	-1 Phishing 1 Legitimate
1.3. HTML and JavaScript based Features 1.3.1. Website Forwarding 1.3.2. Status Bar Customization 1.3.3. Disabling Right Click 1.3.4. Using Pop-up Window 1.3.5. IFrame Redirection	
1.4. Domain based Features 1.4.1. Age of Domain 1.4.2. DNS Record 1.4.3. Website Traffic 1.4.4. PageRank 1.4.5. Google Index 1.4.6. Number of Links Pointing to Page 1.4.7. Statistical-Reports Based Feature	

II. MATERIAL AND METHOD

Procedural steps for solving the classification problem presented is as follows:

- **Identification of the problem**

This study attempts to solve the problem as to how phishing analysis data will be classified.

- **Data set**

Approximately 11,000 data containing the 30 features extracted based on the features of websites in UC Irvine Machine Learning Repository database.

- **Modeling**

After the data is ready to be processed, modeling process for the learning algorithm is initiated. The model is basically the construction of the need for output identified in accordance with the task qualifications.

A. Classification

Classification is to determine the class to which each data sample of the methods belongs, which methods are used when the outputs of input data are qualitative. The purpose is to divide the whole problem space into a certain number of classes. A wide range of classification methods are present. This is due to the fact that different classification methods have been constructed for different data as there is no perfect method that works on every data set. As mentioned in literature studies, the aim of classification is to assign the new samples to classes by using the pre-labeled samples. The most commonly used classification methods are described below.

- Artificial Neural Networks (ANN)
- Support Vector Machine (SVM)
- Naive Bayes (NB)

Extreme Learning Machine (ELM)

Extreme Learning Machine (ELM) is a feed-forward artificial neural network (ANN) model with a single hidden layer. For the ANN to ensure a high-performing learning, parameters such as threshold value, weight and activation function must have the appropriate values for the data system to be modeled. In gradient-based learning approaches, all of these parameters are changed iteratively for appropriate values. Thus, they may be slow and produce low-performing results due to the likelihood of getting stuck in local minima. In ELM Learning Processes, differently from ANN that renews its parameters as gradient-based, input weights are randomly selected while output weights are analytically calculated. As an analytical learning process substantially reduces both the solution time and the likelihood of error value getting stuck in local minima, it increases the performance ratio. In order to activate the cells in the hidden layer of ELM, a linear function as well as non-linear (sigmoid, sinus, Gaussian), non-derivable or discrete activation functions can be used [12-19]. ELM structure is given in Figure 1.

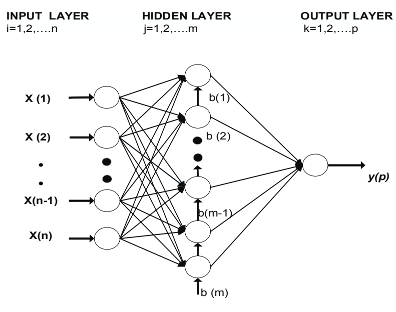

Fig. 1. An artificial neural network model with a single hidden layer with forwardfeed

$$y(p) = \sum_{j=1}^{m} \beta_j \, g\left(\sum_{i=1}^{n} w_{i,j} x_i + b_j\right) \tag{31}$$

In equation 1, xi refers to input vector and yp refers to output vector (m and n neuron count) , wi,j indicates input layer to hidden layer weights and βj indicates output layer to hidden layer weights, bj represents the threshold value of neurons in the hidden layer and g(.) represents activation function. Input layer weights (w) and bias (bj) values in the equation are randomly assigned. Activation function (g(.)), input layer neuron count (n) and hidden layer neuron count (m) are assigned in the beginning step [12-19].

- **Model performance evaluation**

The topics addressed in this section are the two measures that affect the performance of the model and the algorithm used, the first one being the division of data set into training and test data set and the second one being the definition of expressions measuring the performance. In the first measure, the data set is divided into three parts as training, validation and test data by three-phase division in K-Fold method, and model selection and performance status are simultaneously performed. In the second measure, performance assessment of classifier models generally uses a validation value. Validation value can be measured as the ratio of data count detected or estimated correctly by the algorithm into all data in the data set.

$$Accuracy = \frac{A_{poz}+A_{neg}}{Tot} \qquad (32)$$

III. EXPERIMENTAL RESULTS

These results were obtained by using MATLAB 2103b software and a PC with Intel i7-6500 CPU and 8 GB RAM.

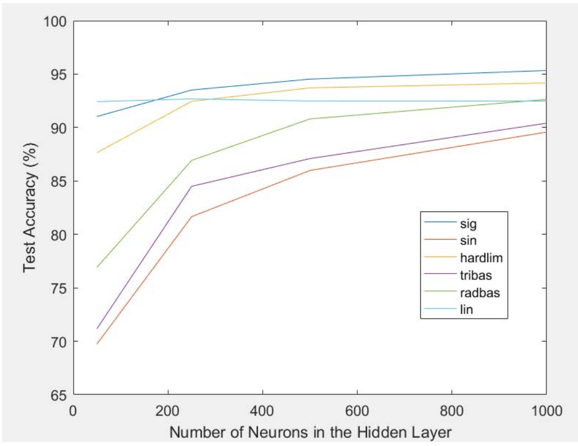

Fig. 2. ELM performance chart.

While attaining these results, cell count in the hidden layer is 1000 and activation count is sigmoid for ELM.

- Comparison of the results of different classification methods

Achieved performance of ELM method and achieved performance of other machine learning methods (Support Vector Machine (SVM), Naive Bayes (NB)) are presented in Table 2. As deduced from these data, ELM achieved higher performance compared to other methods in terms of performance and speed.

TABLE II. ACCURACY OF MACHINE LEARNING METHODS.

Methods	*Train Accuracy*	*Test / True Accuracy*
ELM	**100%**	**95.34%**
NB	100%	93,80%
SVM	100%	92,98%

IV. EXPERIMENTAL RESULTS

In this study, features in the database created for phishing websites are classified by determining the input and output parameters for the ELM classifier. Results obtained by ELM show that ELM has higher achievement compared to other classifier (SVM and NB) methods. This study is considered to be an applicable design in automated systems with high-performing classification against the phishing activity of websites. Furthermore, in literature comparisons, this study is observed to be high-performing by having a high performance of 92.18% that is also the highest test performance in the publication no. [3].

V. CONCLUSIONS

In this paper, we defined features of phishing attack and we proposed a classification model in order to classification of the phishing attacks. This method consists of feature extraction from websites and classification section. In the feature extraction, we have clearly defined rules of phishing feature extraction and these rules have been used for obtaining features. In order to classification of these feature, SVM, NB and ELM were used. In the ELM, 6 different activation functions were used and ELM achieved highest accuracy score.

References

[1] G. Canbek and Ş. Sağıroğlu, “A Review on Information, Information Security and Security Processes,” Politek. Derg., vol. 9, no. 3, pp. 165–174, 2006.

[2] L. McCluskey, F. Thabtah, and R. M. Mohammad, “Intelligent rule-based phishing websites classification,” IET Inf. Secur., vol. 8, no. 3, pp. 153–160, 2014.

[3] R. M. Mohammad, F. Thabtah, and L. McCluskey, “Predicting phishing websites based on self-structuring neural network,” *Neural Comput. Appl.*, vol. 25, no. 2, pp. 443–458, 2014.

[4] R. M. Mohammad, F. Thabtah, and L. McCluskey, “An assessment of features related to phishing websites using an automated technique,” *Internet Technol. …*, pp. 492–497, 2012.

[5] W. Hadi, F. Aburub, and S. Alhawari, “A new fast associative classification algorithm for detecting phishing websites,” Appl. Soft Comput. J., vol. 48, pp. 729–734, 2016.

[6] N. Abdelhamid, “Multi-label rules for phishing classification,” Appl. Comput. Informatics, vol. 11, no. 1, pp. 29–46, 2015.

[7] N. Sanglerdsinlapachai and A. Rungsawang, “Using domain top-page similarity feature in machine learning-based web phishing detection,” in 3rd International Conference on Knowledge Discovery and Data Mining, WKDD 2010, 2010, pp. 187–190.

[8] W. D. Yu, S. Nargundkar, and N. Tiruthani, “A phishing vulnerability analysis of web based systems,” IEEE Symp. Comput. Commun. (ISCC 2008), pp. 326–331, 2008.

[9] P. Ying and D. Xuhua, “Anomaly based web phishing page detection,” in Proceedings - Annual Computer Security Applications Conference, ACSAC, 2006, pp. 381–390.

[10] M. Moghimi and A. Y. Varjani, “New rule-based phishing detection method,” Expert Syst. Appl., vol. 53, pp. 231–242, 2016.

[11] DATASET: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[12] G.-B. Huang et al., “Extreme learning machine: Theory and applications,” Neurocomputing, vol. 70, no. 1–3, pp. 489–501, 2006.

[13] C. S. Guang-bin Huang, Qin-yu Zhu, “Extreme learning machine: A new learning scheme of feedforward neural networks,” Neurocomputing, vol. 70, pp. 489–501, 2006.

[14] T. S. Guzella and W. M. Caminhas, “A review of machine learning approaches to Spam filtering,” Expert Systems with Applications, vol. 36, no. 7. pp. 10206–10222, 2009.

[15] Ö. F.. Ertuğrul, Aşırı Öğrenme Makineleri ile biyolojik sinyallerin gizli kaynaklarına ayrıştırılması. D.Ü. Mühendislik Dergisi Cilt: 7, 1, 3-9-2016

[16] M. E. Tagluk, M. S. Mamiş, M. Arkan, and Ö. F. Ertugrul, “Aşiri Ögrenme Makineleri ile Enerji Iletim Hatlari Ariza Tipi ve Yerinin Tespiti,” in *2015 23rd Signal Processing and Communications Applications Conference, SIU 2015 - Proceedings*, 2015, pp. 1090–1093.

[17] Ö. Faruk Ertuğrul and Y. Kaya, “A detailed analysis on extreme learning machine and novel approaches based on ELM,” Am. J. Comput. Sci. Eng., vol. 1, no. 5, pp. 43–50, 2014.

[18] Ö. F. Ertugrul, “Forecasting electricity load by a novel recurrent extreme learning machines approach,” Int. J. Electr. Power Energy Syst., vol. 78, pp. 429–435, 2016.

[19] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.