A ENHANCED MACHINE LEARNING TECHNIQUE FOR CLASSIFICATION OF PHISHING SITE COMPONENTS

Suresh Naik K[1], Padmavathamma M[2]
[1]PG Student, Department of Computer Science, Sri Venkateshwara University Tirupati
[2]Professor, Department of Computer Science, Sri Venkateshwara University Tirupati

Abstract
Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. This paper discuss about phishing and phishing are a standout amongst the most widely recognized and most perilous assaults among cybercrimes. The point of these assaults is to take the data utilized by people and associations to direct exchanges. Phishing sites contain different indications among their substance and internet browser-based data. The reason for this investigation is to perform Extreme Learning Machine (ELM) based characterization for 30 highlights incorporating Phishing Websites Data in UC Irvine Machine Learning Repository database. For results appraisal, ELM was contrasted and other AI techniques, for example, Support Vector Machine (SVM), Naïve Bayes (NB) and identified to have the most noteworthy exactness.
Keywords: Machine Learning, Support Vector Classifier, Phishing, Information Security.

I. INTRODUCTION

Web use has turned into a fundamental piece of our every day exercises because of quickly developing innovation. Because of this quick development of innovation and serious utilization of advanced frameworks, information security of these frameworks has increased incredible significance. The essential target of keeping up security in data advancements is to guarantee that fundamental safety measures are taken against dangers and threats liable to be looked by clients amid the utilization of these innovations. Phishing is characterized as emulating dependable sites so as to acquire the exclusive data went into sites each day for different purposes, for example, usernames, passwords and citizenship numbers. Phishing sites contain different indications among their substance and internet browser-based data. Individual(s) submitting the misrepresentation sends the phony site or email data to the objective location as though it originates from an association, bank or whatever other dependable source that performs solid exchanges. Substance of the site or the email incorporate solicitations planning to draw the people to enter or refresh their own data or to change their passwords just as connections to sites that resemble precise of the sites of the associations concerned. 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.

II RELATED WORK

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.

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)

III PROPOSED SYSTEM

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.

IV METHODOLOGY

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.

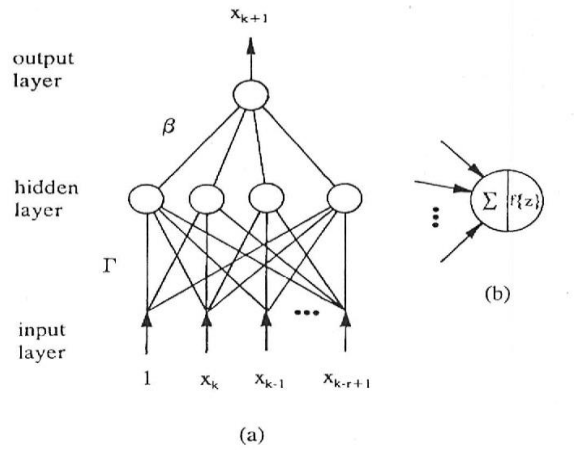

Fig: An Artificial Neural Network Model

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.

V CONCLUSION

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.

VI REFERENCES

[1] G. Canbek and ù. Sa÷Õro÷lu, "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.. Ertu÷rul, AúÕrÕ Ö÷renme Makineleri ile biyolojik sinyallerin gizli kaynaklarÕna ayrÕútÕrÕlmasÕ. D.Ü. Mühendislik Dergisi Cilt: 7, 1, 3-9- 2016 [16] M. E. Tagluk, M. S. Mamiú, M. Arkan, and Ö. F. Ertugrul, "Aúiri 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 Ertu÷rul 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.

K SURESH NAIK he is a master of Computer Science (M.Sc) pursuing in Sri Venkateswara University, Tirupati, A.P. He received Degree of Bachelor of Science in 2017 from Rayalaseema University, Kurnool. His research interests are Cloud Computing, Artificial Intelligence, and Big Data.