International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Detection of Phishing Web Sites Based On Feature Classification and Extreme Learning Machine"

Mr.Pankaja Kumar Kandi

Department of Computer Engineering, DYPSOE, Pune

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Phishing locales which hopes to take the exploited people private information by occupying them to surf a phony site page that looks like a true blue one is another kind of criminal acts through the web and its one of the particularly worries toward various territories including e-dealing with a record and retailing. Phishing site identification is genuinely a capricious and component issue including various parts and criteria that are not steady. We proposed an insightful model for distinguishing phishing website pages dependent on Extreme Learning Machine. Sorts of site pages are diverse as far as their highlights. Thus, we should utilize a particular website page highlights set to forestall phishing assaults. We proposed a model dependent on AI methods to distinguish phishing website pages. We have recommended some new guidelines to have effective element grouping. The model has 30 data sources and 1 yield. Right now, 10-overlap cross-approval test has been performed. The normal grouping precision estimated as 95.05%.

Keywords*: Phishing, Extreme Learning Machine, Feature Classification.*

Introduction

Innovation is developing quickly step by step and with this fast developing innovation web has become a fundamental piece of human's day by day exercises. Utilization of web has become because of the fast development of innovation and serious utilization of advanced frameworks and in this way information security has increased incredible significance. The essential goal of keeping up security in data advances is to guarantee that fundamental safety measures are taken against dangers and risks liable to be looked by clients during the utilization of these advances. Phishing is the deceitful endeavor to acquire touchy data, for example, usernames, passwords and Mastercard subtleties by masking as a reliable element in an electronic correspondence. Ordinarily did by email parodying or texting, it regularly guides clients to enter individual data at a phony site, the look and feel of which is indistinguishable from the real site. Information security threats have been seen and developed through time along development in the internet and information systems. The impact is the intrusion of information security through the compromise of private data, and the victims may lose money or other kinds of assets at the end. Internet users can be affected from different types of cyber threats such as private information loss, identity theft, and financial damages. Hence, using of the internet may suspect for home and official environments.

Identify and defend against privacy leakage efficient analytical tools are required for users to reduce security threats. Effective systems that can improve self-intervention must be formed using artificial intelligence-based information security management system at the time of an attack. Phishing is an Internet-based attack that seduces end users to visit fake websites and give away personal information such as user id and password. Phishing web pages are formed by fraudulent people to copy a web page from an original one. These phishing web pages are very similar to the original ones. Technical tricks and social engineering are extensively joined together for beginning a phishing attack. An important view of online security is to protect users from phishing attacks and fake websites. Intelligent methods can be used to develop fake web pages. For this reason internet users whether have enough experience in information security or not might be cheated. Phishing attacks can be launched via sending an e-mail that seems to be sent from a trusted public or private organization to users by attackers. Attackers get the users to update or verification their information by clicking a link within the e-mail. Other methods such as file sharing, blogs, and forums can be used by attackers for phishing. There are many ways to fight phishing including legal solutions, education, and technical solution. A significant number of studies on the phishing have been done.

Nowadays, information and communication tools are used in a manner that is very dense with information. For this purpose, various solution methods for various problem types have been developed. Machine Learning (ML) methods, can also be used in application development for information security. Optimization, classification, prediction and decision support system and great benefits can be provided to the person who is responsible for information security. Today, it has become an increasingly popular subject in developing intelligent applications. Non-intelligent application can cause losses in case the user is not required and can do a job that requires again.

There are attacks for different purposes to the Information and Communication tools that create computer networks. These attacks can be detected and the necessary precautions should be taken. For the study of artificial intelligence seems to gain speed as computer technology evolves. Artificial knowledge strategies and concentrates on data security are expanding step by step. Canny frameworks give extraordinary advantages in choosing to data security experts.

ML strategies can be utilized with arrangement purposes in different fields. Arrangement can be considered as a process to determine whether a data belong to one of the classes in the dataset organized according to certain rules. Classification which used in many fields and has an important place has a separate place for information security.

Review of Literature

Santhana Lakshmi and Vijaya used Machine-learning technique for modeling the prediction task and supervised learning algorithmsthat Multi-LayerPerceptron. Decision tree and Naïve bayes classifications were used for observing. It has been observed that the decision tree classifier predicts the phishing website more accurately then other learning algorithms.

Zou Futai, Pei Bei and Panli proposed Uses Graph Mining technique for web Phishing Detection. It can identify some potential phishing which can't be distinguished by URL investigation. It uses the meeting connection among client and site. To get dataset from the genuine traffic of a Large ISP. After anonym punch these information, they have purifying dataset and each record incorporates eight fields: User hub number (AD), User SRC IP (SRC-IP) get to time (TS), Visiting (URL), Reference URL (REF), User Agent (UA), get to server IP (DSTIP), User (treat). Xin Mei Choo, Kang Leng Chiew proposed a strategy which concentrate and structure the list of capabilities for a page. It utilizes a SVM machine as a classifier which has two stages preparing stage and testing stage during preparing stage it extricates highlight set and keeping in mind that testing it foresee the site is genuine or a phishing.

Kaytan and Hanbay proposed deciding phishing sites dependent on neural system. UCI (University of California, Irvine) dataset was utilized for the examination. 30 information traits, and 1 yield property were utilized for the trial. The qualities 1, 0, and - 1 were utilized for input characteristics and the qualities 1, and - 1 were utilized for yield property.5-fold cross validation method was used for evaluating the system performance. The best classification accuracy has been measured as 92.45%. And the average accuracy has been measured as 90.61%.

Chen et al evaluated intensity of phishing attacks in terms of risk levels and potential market value losses experienced by the target companies. It was analyzed 1030 phishing alerts released on a public database, and financial data related to the targeted firms using a hybrid method. The severity of the attack was predicted with up to 89% accuracy using text phrase extraction and supervised classification. It has been identified some important textual and financial variables in the study. Impact the severity of the attacks and potential financial loss has been investigated.

Giovanni Armano and Samuel Marchal proposed a novel approach based on minimum enclosing ball support vector machine (BVM) to detect phishing website. It has been aimed at achieving high speed and high accuracy to detect phishing website. Studies were done in order to enhance the integrity of the feature vectors. Firstly, an analysis of the topology structure of website was performed according to the Document Object Model (DOM) tree. Then, the web crawler was used to extract 12 topological features of the website. Later, the feature vectors were detected by BVM classifier. The proposed method was compared to the DVM. It was observed that the proposed method has relatively high precision of detecting. In addition, it was observed that the proposed method complements the disadvantage of slow speed of convergence on large-scale data. It has been shown that the proposed method has better performance than SVM in the experimental results. The accuracy and validity of the proposed system has been evaluated.

Gowtham and Krishnamurthi studied the characteristics of legitimate and phishing web pages in depth. Heuristics were proposed to extract 15 features from similar types of web pages based on the analysis. The proposed heuristic results were fed as an input to a trained machine learning algorithm to detect phishing websites. Before the applying the heuristics to the site pages, two starter screening modules were utilized in the framework. By the preapproved website identifier that is the principal module, site pages were checked against a private whitelist kept up by the client. By the login structure discoverer that is the subsequent module, site pages were delegated genuine when no login structures present. Superfluous calculation in the framework was decreased by helping the pre-owned modules. Furthermore, the pace of

bogus positives without settling on the bogus negatives was decreased by helping the pre-owned modules. By utilizing the modules, website pages have been ordered with 99.8% accuracy and a 0.4% of bogus positive rate. It has been demonstrated that the proposed technique is proficient for shielding clients from online character assaults. The principal subject is about the calculation of expected limits to depict the three email gatherings. Furthermore, the subsequent point is the translation of the cost-touchy attributes of spam separating. They reliably figure the choice theoretic harsh set model based edges. The blunder pace of misclassification an authentic email to spam is watched. What's more, it has been seen that the new technique diminishes the blunder rate. The examination speaks to a superior presentation so as to the costaffectability viewpoint.

Proposed System

The proposed methodology which imports dataset of phishing and legitimate URLs from the database and the imported data is preprocessed. Detecting phishing website is performed based on four categories of URL features: domain based, address based, abnormal based and HTML, JavaScript features. These URL features are extracted with processed data and values for each URL attribute are generated. The analysis of URL is performed by machine learning technique which computes range value and the threshold value for URL attributes. Then it is classified into phishing and legitimate URL. The attribute values are computed using feature extraction of phishing websites and it is used to identify the range value and threshold value. The value for each phishing attribute is ranging from {-1, 0, 1} these values are defined as low, medium and high according to phishing website feature. The classification of phishing and legitimate website is based on the values of attributes extracted using four types of phishing categories and a machine learning approach.

URL Feature Analysis

The phishing attribute features are extracted for each URL to find whether the website is phishing or legitimate. The URL_of_Anchor tag attribute is selected to find the overlap values. The overlap value is the sum of selected attribute value which is combined with other attributes.

Finding Attribute Values

The attribute value for each URL is computed using corresponding set of attribute values {- 1, 0, 1}. Fig 1 represents attribute X that URL_of_Anchor tag value and attribute Y that is Prefix_Suffix value. Both the attributes URL_of_Anchor tag and Prefix_Suffix also have inter linked value and that has to be computed for finding range and threshold value.

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.

System Architecture

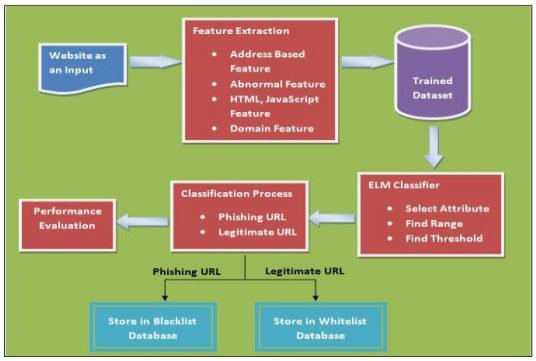
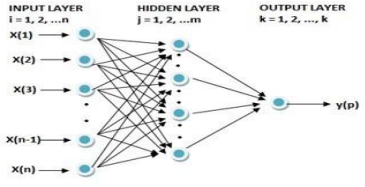


Figure 1: System architecture

Algorithm

Extreme Learning Machine (ELM)

Extraordinary Learning Machine (ELM) is proposed as a solitary concealed layer feed-forward counterfeit neural system (ANN) model which guarantee a high-performing learning and parameters, for example, limit worth, weight and actuation the capacity must have fitting qualities for the information framework which is to be displayed. In ELM learning, the parameters are angle based, where the information loads are haphazardly chosen while the yield loads are scientifically determined. For enacting the cells in the shrouded layer of ELM, a direct capacity just as non-straight (sinus, sigmoid, Gaussian), and the nonlogical or discrete actuation capacities can be utilized.



Here, n: training samples, m: number of classes, i = 1, 2,..., n, j = 1, 2,...m, k = 1, 2,..., k, xi: input vector and y(p): desired output vector.

There are three layers, which are input layer, the hidden layer and the output layer.

$$y(p) = \sum_{j=1}^{m} \beta j \, a(\sum_{i=1}^{n} wi,j.xi + bj) \quad .....(1)$$

In above equation 1, wi,j is an input layer to hidden layer weights and βj is an output layer to hidden layer weights, bj is the threshold value of neurons in the hidden layer and a(.) is the activation function. In the input layer, weights (w) and bias (bj) values are randomly assigned in the equation. The activation function (a(.)), input layer neuron count (n) and hidden layer neuron count (m) are
assigned in the beginning
Step 1: Enter a URL of a website.
Step 2: Examine all the attributes of the website or the web page according to its features.
Step 3: Fetch all the samples features to the dataset.
Step 4: Randomly select 10% of the testing samples while 90% training samples of the dataset.
Step 5: Apply ELM classification algorithm on the dataset Step 5.1: Arbitrarily generate hidden node parameters.
Step 5.2: Calculate output matrix for the hidden layer.
Step 5.3: Calculate weight of the output matrix. Step 6: Prediction for website whether phishing or legitimate.

System Requirements

A. Software Requirement

1. Operating System: Windows 7 or above
2. Programming Language: Python 3.7
3. IDE: Python IDLE

B. Hardware Requirement

1. Processor: Pentium Processor Core 2 Duo or Higher
2. Hard Disk: 250 GB (min)
3. RAM: 1GB or higher
4. Processor Speed: 3.2 GHz or faster processor

Experimental Analysis

The result obtained by ELM classifier has greater accuracy achievement as compared to the other classifiers i.e. Support Vector Machine (SVM) and Naive Bayes (NB) methods. The study is thus considered to be an applicable design with high performing classification against the hazardous phishing activity of the websites. Also, if we compare the literature study the proposed study is observed to be high-performing this has greater accuracy of 92.18% which is also the highest accuracy in the publication.

Classification Method	TrainAccuracy	TestAccuracy
Extreme Machine Learning (ELM)	100%	96.93%
Support Vector Machine (SVM)	100%	94.80%
Naive Bayes (NB)	100%	54.38

Fig. 2: Accuracy Output

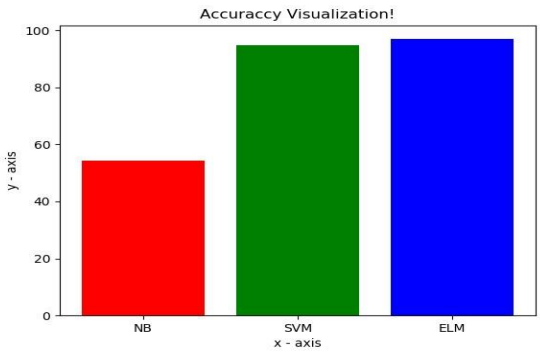


Fig. 3: Accuracy Visualization

Conclusion

Systems varying from data entry to information processing applications can be made through websites. The entered information can be processed; the processed information can be obtained as output. Nowadays, web sites are used in many fields such as scientific, technical, business, education, economy, etc. Because of this intensive use, it can be also used as a tool by hackers for malicious purposes. One of the malicious purposes emerges as a phishing attack. A website or a web page can be imitated by phishing attacks and using various methods. Some information such as users credit card information, identity information can be obtained with these fake websites or the web pages. The purpose of the application is to make a classification for the determination of one of the types of attacks that cyber threats called phishing. Extreme Learning Machine is used for this purpose. In this study, we used a data set taken from UCI website. In this dataset, input attributes are determined in 30, and the output attribute is determined in 1.Input attributes can take 3 different values which are 1, 0, and -1.Output attribute can take 2 different values which are 1, and -1. As a result of the study, the

average classification accuracy was measured is 96.93%.

References

[1]. N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41(13), pp. 5948-5959, 2014.

[2]. R. M. Mohammad, F. Thabtah, L. McCluskey, "Tutorial and critical analysis of phishing websites methods," Computer Science Review, vol.17, pp. 1-24, 2015.

[3]. R. M. Mohammad, F. Thabtah, L. McCluskey, "Predicting phishing websites based on selfstructuring neural network," Neural Computing and Applications, vol. 25(2), pp. 443-458, 2014.

[4]. M. A. U. H. Tahir, S. Asghar, A. Zafar, S. Gillani, "A Hybrid Model to Detect PhishingSites Using Supervised Learning Algorithms," International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1126-1133, IEEE, 2016.

[5]. R. M. Mohammad, F. Thabtah, L. McCluskey, "Intelligent Rule-based Phishing Websites Classification, " IET Information Security, vol. 8(3),pp. 153-160, 2014.

[6]. Hodzic, J. Kevric, A. Karadag, "Comparison of Machine Learning Techniques in Phishing Website Classification," 2016.