EasyChair Preprint

№ 2425

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

Pankaj Kumar Kandi and Pankaj Agarkar

January 20, 2020

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

Mr.Pankaja Kumar Kandi
Department of Computer Engineering, Pune
Savitribai Phule Pune University
Pune, India
pkkandi@gmail.com

Prof. Pankaj Agarkar
Department of Computer Engineering, Pune
Savitribai Phule Pune University
Pune, India
pankaj.agarkar@dypic.in

ABSTRACT: Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. We proposed an intelligent model for detecting phishing web pages based on Extreme Learning Machine. Types of web pages are different in terms of their features. Hence, we must use a specific web page features set to prevent phishing attacks. We proposed a model based on machine learning techniques to detect phishing web pages. We have suggested some new rules to have efficient feature classification. The model has 30 inputs and 1 output. In this application, the 10-fold cross-validation test has been performed. The average classification accuracy measured as 95.05%.
KEYWORDS: Phishing, Extreme Learning Machine, Feature Classification.

I. INTRODUCTION

Technology is growing rapidly day-by-day and with this rapid growing technology internet has become an essential part of human's daily activities. Use of internet has grown due to the rapid growth of technology and intensive use of digital systems and thus data security 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. Phishing is the fraudulent attempt to obtain sensitive information such as usernames, passwords and credit card details by disguising as a trustworthy entity in an electronic communication. Typically carried out by email spoofing or instant messaging, it often directs users to enter personal information at a fake website, the look and feel of which is identical to the legitimate 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 intelligence methods and studies on information security are increasing day by day. Intelligent systems provide great benefits in deciding to information security professionals.

ML methods can be used with classification purposes in various fields. Classification 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.

II. REVIEW OF LITERATURE

Santhana Lakshmi and Vijaya used Machine-learning technique for modeling the prediction task and supervised learning algorithms that Multi-Layer Perceptron. 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 detect some potential phishing which can't be detected by URL analysis. It utilizes the visiting relation between user and website. To get dataset from the real traffic of a Large ISP. After anonym zing these data, they have cleansing dataset and each record includes eight fields: User node number (AD), User SRC IP (SRC-IP) access time (TS), Visiting URL (URL), Reference URL (REF), User Agent (UA), access server IP (DSTIP), User cookie (cookie).

Xin Mei Choo, Kang Leng Chiew proposed a method which extract and form the feature set for a webpage. It uses a SVM machine as a classifier which has two phases training phase and testing phase during training phase it extracts feature set and while testing it predict the website is legitimate or a phishing.

Kaytan and Hanbay proposed determining phishing websites based on neural network. UCI (University of California, Irvine) dataset was used for the study. 30 input attributes, and 1 output attribute were used for the experiment. The values 1, 0, and -1 were used for input attributes and the values 1, and -1 were used for output attribute. 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 web pages, two preliminary screening modules were used in the system. By the preapproved site identifier that is the first module, web pages were checked against a private white-list maintained by the user. By the login form finder that is the second module, web pages were classified as legitimate when no login forms present. Unnecessary computation in the system was reduced by helping the used modules. Additionally, the rate of false positives without compromising on the false negatives was reduced by helping the used modules. By using the modules, web pages have been classified with 99.8% precision and a 0.4% of false positive rate. It has been shown that the proposed method is efficient for protecting users from online identity attacks. The first topic is about the computation of required thresholds to describe the three email groups. And the second topic is the interpretation of the cost-sensitive characteristics of spam filtering. They consistently calculate the decision-theoretic rough set model based thresholds. The error rate of misclassification a legitimate email to spam is observed. And it has been seen that the new method reduces the error rate. The study represents a better performance in order to the cost-sensitivity perspective.

III. PROPOSED SYSTEM

The proposed methodology which imports dataset of phishing and legitimate URLs from the database and the imported data is pre-processed. 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.

IV. SYSTEM ARCHITECTURE

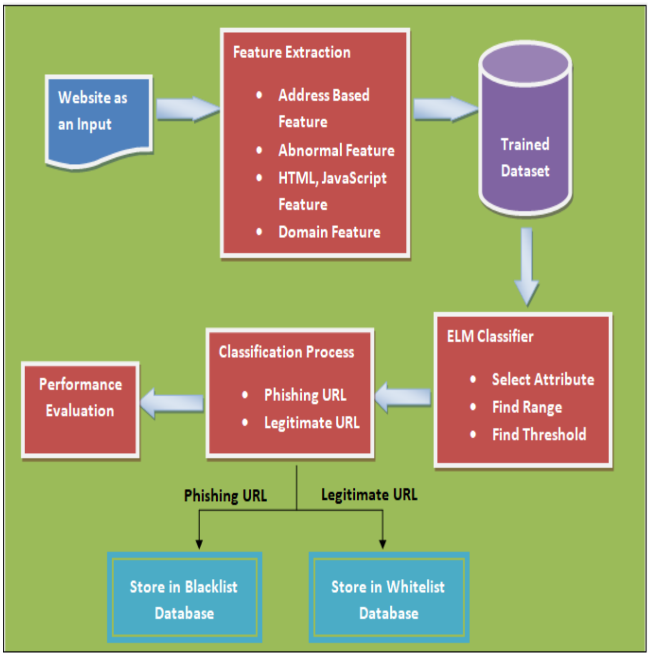
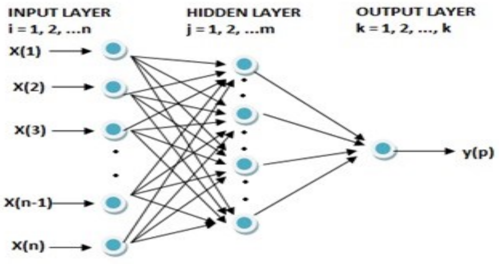


Figure 1: System architecture

V. ALGORITHM

Extreme Learning Machine (ELM)
Extreme Learning Machine (ELM) is proposed as a single hidden layer feed-forward artificial neural network (ANN) model which ensure a high-performing learning and parameters such as threshold value, weight and activation the function must have appropriate values for the data system which is to be modeled. In ELM learning, the parameters are gradient-based, where the input weights are randomly selected while the output weights are analytically calculated. For the sake of activating the cells in the hidden layer of ELM, a linear function as well as non-linear (sinus, sigmoid, Gaussian), and the non-derivable or discrete activation functions can be used.



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.

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

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.

VI. SYSTEM REQUIREMENTS

A. Software Requirement

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

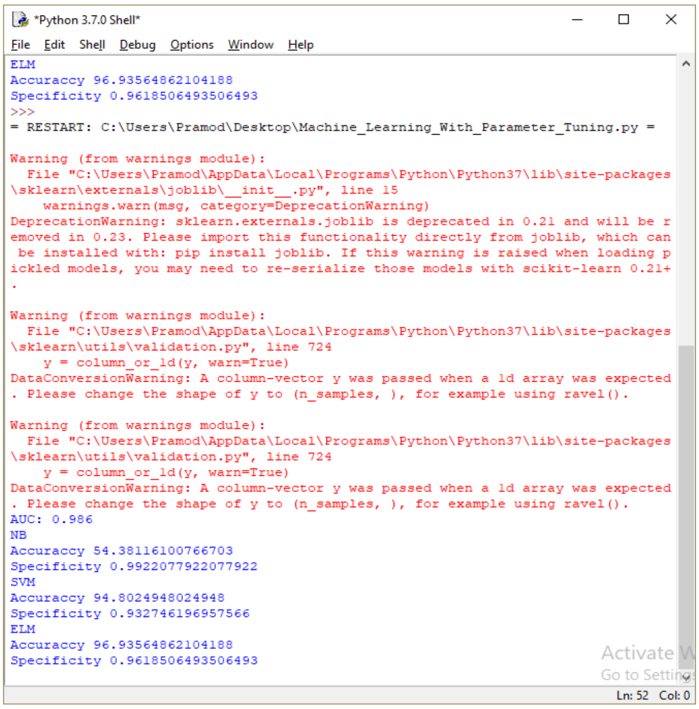
B. Hardware Requirement

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

VII. 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	Train Accuracy	Test Accuracy
Extreme Machine Learning (ELM)	100%	96.93%
Support Vector Machine (SVM)	100%	94.80%
Naive Bayes (NB)	100%	54.38



```
ELM
Accuraccy 96.93564862104188
Specificity 0.9618506493506493
>>>
= RESTART: C:\Users\Pramod\Desktop\Machine_Learning_With_Parameter_Tuning.py =

Warning (from warnings module):
  File "C:\Users\Pramod\AppData\Local\Programs\Python\Python37\lib\site-packages
\sklearn\externals\joblib\__init__.py", line 15
    warnings.warn(msg, category=DeprecationWarning)
DeprecationWarning: sklearn.externals.joblib is deprecated in 0.21 and will be r
emoved in 0.23. Please import this functionality directly from joblib, which can
 be installed with: pip install joblib. If this warning is raised when loading p
ickled models, you may need to re-serialize those models with scikit-learn 0.21+
.

Warning (from warnings module):
  File "C:\Users\Pramod\AppData\Local\Programs\Python\Python37\lib\site-packages
\sklearn\utils\validation.py", line 724
    y = column_or_1d(y, warn=True)
DataConversionWarning: A column-vector y was passed when a 1d array was expected
. Please change the shape of y to (n_samples, ), for example using ravel().

Warning (from warnings module):
  File "C:\Users\Pramod\AppData\Local\Programs\Python\Python37\lib\site-packages
\sklearn\utils\validation.py", line 724
    y = column_or_1d(y, warn=True)
DataConversionWarning: A column-vector y was passed when a 1d array was expected
. Please change the shape of y to (n_samples, ), for example using ravel().
AUC: 0.986
NB
Accuraccy 54.38116100766703
Specificity 0.9922077922077922
SVM
Accuraccy 94.8024948024948
Specificity 0.932746196957566
ELM
Accuraccy 96.93564862104188
Specificity 0.9618506493506493
```

Fig. 2: Accuracy Output

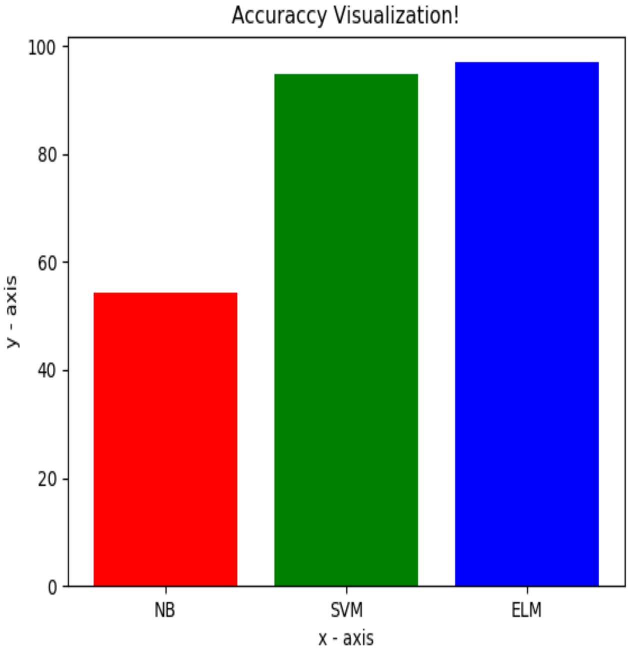


Fig. 3: Accuracy Visualization

VIII. 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 self-structuring 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 Phishing-Sites 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.