Phishing Detection using Multi-Layer Perceptron and Comparison of Accuracy with Various Neural Network Techniques

Akshay Hanmantrao Thotwe[1] Dr. Sunil B. Mane[2]
[1,2]Department of Computer Engineering [2]Department of Computer Engineering & Information Technology
[1,2]College of Engineering, Pune(COEP), Shivajinagar, Pune-411005, India

Abstract— In this modern world number of users connected to internet are increasing rapidly with respect to number of attacks on those user to steal an information. The user data is compromised by using various hacking method implemented by hacker on system. In this paper, we will study Phishing Detection using Multi-Layer Perceptron and make comparison of accuracy obtained by using various solving and with respect to activation function. In this model, we have built Neural Network with the help of sklearn. The Dataset used to train the Neural Network has been taken from UCI machine Learning Repository for Phishing web set dataset. The Dataset contains both phishing and non-phishing instances. The number of Phishing and Non-phishing are 4898 and 6157 instances. Instead of feature selection we have extracted those feature that are important to decide legitimacy of the website. We also considered the time taken to build the neural model. After classifying the instances into phishing and legitimate, we got high accuracy. Also we have used Sklearn, it has different solving and activation technique that are defined in Multi-Layer Perceptron. We are classifying solving methods based on accuracy. Also, we have used weka tool to check the accuracy of all neural network algorithm defined in it. In this, we have classified new instances into phishing and non-phishing by using multi-layer perceptron then compared its accuracy with the various neural network techniques.

Key words: Phishing Detection, Artificial Neural Network, Multi-Layer Perceptron, Sklearn, Backpropogation, Feed-Forward Neural Network, Voted Perceptron, Weka Tool, Machine Learning

I. INTRODUCTION

Todays Modern World the availability of internet has increased with respect to number of users connected to it. There are number of contents and web sites present on internet. The user access those database continuously. So, the number of users access a popular website is also increasing. According to study, the number of users connected to copy of that website are increasing, because the hackers created the web page which exactly looks like a legitimate website. The World-Wide-Web is filled with such sites. The user are unable to distinguish between the fake and original website until he analyses the HTML page of that website. Due to lack of time to study about the website user continue using that site, hence results to compromising information of that user in the hands of hacker. Phishers are creating the web pages that exactly looks like the given website even uses different techniques to attract users. Phishing has caused lots of loss to financial sector. So a Phishing Detection is important to prevent the user information, login credentials, and bank details. The main aim of designing phishing website detection model is that should able to classify any web sites even they are newly created. The motivation behind the project that increasing use of artificial intelligence in every corner, so tried to design the intelligent model to classify the phishing websites.

The various reserchers used WEKA tool to analyze the effictiveness of the neural network model over the other machine learning algorithm in terms of accuracy of model. In this survey found that training and testing on dataset model used and stated accuaracy. But we are generalising the model by extracting feature of selected web site and providing it to model and get validity of the website i.e. whether it is phishing or not.

The rest of the paper is structured as follows: Section 2 for literature Survey, Section 3 for Proposed Methodology, and Section 4 for results and it is followed by section 5 which is composed of conclusion

II. LITERATURE SURVEY

A lot of work proposed to design phishing websites, Most of them concentrated on URLs or contents of source code of that site. Some of them used URLs and source code of that site. While designing this model we have considered the importance of both the contents URL and source code of that site. This section of contains the study of related works which helped us to design the model.

1) In this Study, authors proposed classification of various machine learning algothims on basis of supervised machine learning. It is the search for algorithms that reason from extenally supplied instances to produce general hypotheses, which then make predictions about future instances. It also explains about goal of supervised learning is to build a model of the distributed class labels in terms of predictor feature. Then the resulting classifier is then used to assign class labels to testing the instances where the values of the in-put are known but class label are known. It also states the difference interms of accuracy and tolerance towards various states.
2) In this study, authors proposed model which is based on neuron-fizzy. It is rule based model which fuzzy inferences system which is combined with neural network for classification. The Fuzzy logic is used to make predictions and generates fuzzy model based on fuzzy rules.
3) In this study, authors proposed a brief explanation about Multilayer Perceptron model and techniques for its implementation in real world. It states the drawback of the system and its advantage.
4) In this study, author's gives explanation about the supervised machine learning is essential for learning artificial neural network. This model gives the comparitive study between the back propogation algorithms, decision tree algorithm. It evaluates their

efficiency in the range of limited parameters like speed learning, overfitting avoidance and their accuracy

5) In this study, author experimentally proves that previously the multi-layer perceptron have zero output and used use separate short connections to model the linear dependencies. This paper showed that usefulness of transformation by making it stochastic gradient learning, a state of learning in algothims in speed.
6) In this study, authors proposed a system where various machine learning algorithm can be designed for medium scale supervised and un-super-vised problems. In this author focused on bringing machine learning to no specialists by using general purpose high-level learning.
7) In this study, authors have classified the webpage phishing or not on the basis of identity keyword, extraction of keyword and target domain name finder whether they are legitimate or not. This helps in feature extraction and analyzing keyword where they appear in webpage which helped to classify that webpage.
8) In this study, author explains a method for detection based on machine learning classifier along with the wrapper function. Wrapper function is the helped to selection of efficient and significant feature to predict phishing websites accurately. This study helps to state all the feature which are extracted not help to predict the phishing we site.
9) In this study, author describes and investigate the feature selection which helps achieve goal to determine the effective set of feature in terms of classifying performance of the system. It compare two existing feature selection method in order to determine least set of feature for phishing detection in data mining.
10) In this study, author illustrates classification of phishing email by extracting 23 keywords from email bodies and it selects 12 features by using t-statistics. This model classifies accuracy achieved using different machine learning algorithm. It states that Multi-Layer Perceptron have the 97.2respectively. Also the accuracy, sensitivity and specificity with feature selections are 96.72
11) In this study, author designed model for detection of redirection spam using Multilayer Perceptron using neural network. It gives idea about how the proposed methodlogy can be implemented while determining the phishing websites. This model gives 99.28
12) In this study, author states different machine learning algorithm for modeling the prediction task and supervised machine learning algorithm. It classifies model on the basis of their accuracy, time required to build model training and other evaluation criteria in this technique.

III. PROPOSED METHODOLOGY

In this section we will discuss the steps that we followed while designing the model. This section constist of the dataset construction, Feature extraction, various neural network algorithm.

A. Dataset Construction

In This model we have used dataset from UCI Machine learning algorithm for phishing website detection. The dataset consists of 11055 instances. In that dataset there are 4898 Phishing website instances and the 6157 legitimate website instances. In the study we have found that if we used feature selection on these 31 features and choose the efficient feature. This will leads to increasing the training time and decreasing the accuracy of Multi-Layer Perceptron. Due to this we have selected those feature that are useful to found phishing websites. After study on each concept we have selected 21 features. These features are filtered amoung with the result and we formed new dataset based on this selected features. Even if the dataset is old but consists of the correct predictions, when we feed it to neural model then we will get correct result. In this model the dataset is important because of we train the model using wrong dataset then it will give the wrong result. The problem for detecting phishing website it needs to have latest updated dataset but the neural network can make prediction using old dataset.

B. Feature Selection

In this model instead of using the various feature selection method. We have extracted only those feature in our study we have found them important. Features are important for detecting phishing websites. So in our study, we have analysed URL, source code of the site and checked the database of WHO-Is database to parse the given site. The Features that we have considered important are as follows:

1) URL having IP address in domain: If the URL is having IP address as its domain. Then it is considered as phishing we site because name of the domain is hidden under IP. If the IP present we consider it as phishing otherwise we consider it as non-phishing (or legitimate).
2) URL has long Length: Malicious URL has long URL because it has hidden URL in iframe it will expand when one clicks the link. We consider Length of URL determine validity of URL. If the length of URL is less 54 then we consider it as legitimate. If its length is between 54 and 75, then the URL is considered as suspicious. Else if the length of URL is greater than 75, then it is lebeled as phishing.
3) Use of Tiny URL instead of original URL: Hackers generate the tiny-url of the malicious link and send it to the user. User will not understand the link is original or not or the link hidden under it. We consider if the tiny url is present instead of original URL, then we consider it is as phishing website. Othwise, we consider the given site is legitimate.
4) If @; sysmbol present in the URL: Phishers add @ url in the URL because the characters after @ symbol are ignored by the browser. The user leads to conclusion that it is different from the actual nature of the site. We have considered that if the URL has @ sysmbol we consider it as phishing otherwise it is considered as legitimate.
5) If // present in the URL: In the URL // present then it redirect the URL to different sites. We do not parse the redirected website. So, if // present in the URL we consider it as phishing otherwise considered as a legitimate.
6) If - present in the domain name part: Many users mistakely considers the website is legitimate when '-' present in the domain name of the URL. If the - in

domain part is considered as phishing, otherwise it is considered as legitimate.

7) Multiple sub-domains in the URL: If there are multiple sub-domains present in the URL. In this we ignore www part of the URL even if it is considered as sub-domain. If the URL contains one sub-domain then it considered as legitimate. If the number of sub-domain are 2 then it called as suspicious. Otherwise considered as phishing.
8) HTTPS weith Secure Sockets Layer (SSL): HTTPS tag gives assurity of websites legitimacy but not enough criteria. In this feature we check that the HTTPS token and validity of SSL certificate. If both are valid then it is termed as Legitimate. If HTTPS assigned with invalid certificate then termed as Suspicious. Otherwise, it is termed as Phishing.
9) Domain Registration Length: Phishing website lived for short period of time. Legitimates sites use domains are paid of several years in advance. Phishing sites registered for one year only. If the sites expiry date is more than the one or more year termed as Legitimate.
10) Favicon: A favicon is a graphic image assigned with specific webpage. If a favicon is loaded from domain name other than address bar. If the website contains favicon loaded domain considered as Phishing. Otherwise considered as legitimate.
11) Using Non-Standard Port: Firewalls, Proxy and Network Address Translation (NAT) servers block most of the port and only open those port which are selected. Phishers run almost any service they want if all the port are kept open, then phishers try to steal the information. If the port is not prefered status that is OPEN or CLOSE. If the status is wrong then it is termed as phishing. Otherwise it is called as legitiamate
12) HTTPS token in domain name: If the https token pesent in the domain name then it is called as phishing otherwise termed as legitimate.
13) Request URL: This features examine the external objects contained in the webpage such as Images, Videos and Audio shares domain names as the url domain ot not. If the number URLs is less than 17 then it is legitimate site. If it is between 22 and 61 then it is Suspicious. Otherwise it is Phishing.
14) URL of Anchor: The number of links contained in the webpage contains the same domain name as the URL or not, also whether it link to any page or not. These information are extracted using this features. If the URL of anchor is less 12than 31, then it is Legitimate.If it is between 31 to 67, then it is suspicious. Otherwise labeled as Phishing.
15) Submitting information to Email: If mail() or mailto used to redirect users information to attackers personal mail or any other email. If these values present in web form on server side script then it is labeled as phishing. Otherwise labeled as legitimate.
16) Abnormal URL: For this feature, we examine WHOIS database. If the hostname is included in database then it is labeled as legitimate. Otherwise labeled as phishing.
17) Iframe: Iframe a HTML tag used to display as additional webpage. If the iframe tag is found without frame borders then labeled as phishing. Otherwise labeled as legitimate.
18) Age of Domain: For this feature we examine WHOIS database. If the age of domain is less than 6 months then labeled as phishing. Otherwise labeled as legitimate.
19) DNS Record: We examine DNS record of website whether it is present or not in database. If their is no DNS record for the website is present then it is labeled as Phishing. Otherwise labeled as legitimate.
20) Website Traffic: This features examine popularity of website by checking number of visitors visited the website and its pages. We have found that if the website rank is among 1,00,000 then labeled as legitimate. If it is greater then 1,00,000 then it is labeled as suspicious. Otherwise labeled as phishing.

C. Artificial Neural Network Model

Artificial Neuaral Network computational model. It works similar to computation of information by neurons present in human brain. We have used the supervised machine learning algorithm. So the supervised machine learning algorithm of neural network are Multi-layer Percerptron, Feed-Forward Neural Network, and Voted Perceptron. In this section we will briefly explain each algorithm that are mentioned above.

1) Single Layer Perceptron

The single layer perceptron consists one hidden layer where the input neurons are connected to neurons from hidden layer then they are connected to output nodes. The single layer perceptron only capable of linearly separable model. It failed to detect the instances that are present on the both region marked axis of the model. This model consists of threshold value 0 or 1. The activation function used in this model is sigmoid function. This model consist of only one hidden layer so the loss per epoch is minimum but the whole accuracy of this model reduced due to the un separable result.

2) Voted Perceptron

The voted perceptron is model that defined in the WEKA TOOL, we have used this only to compare the accuracies of different supervised neural algorithm. Voted Perceptron is recurrance model used to make prediction by computing prediction algorithm, The voted percptron is fast learning algorithm but the accuracy obtained by using this model is less as compared to other techniques.

3) Multi- Layer Perceptron

This neural model is just like the single layer perceptron which consists of two or more hidden layer sizes. The hidden layer sized can be calulated by using formulae to get maximum stable model to avoid underfitting, overfitting and reducing number of error per epoch or at hidden neuron output calculation. The connection is present between the input node to hidden node and hidden node to output node. The connection present in the backpropagation algorithm are also present between input and the output node. The backpropagation model increase accuracy of computing model by making it fault tolerent. The multi-Layer Perceptron model can detect the model which is present not separable region and classifies them. Thats why it is selected as efficient technique among the other three technique. The threshold are given to calculating the output of each node are 0 or 1, the weight of node helped to calculate the output of hidden nodes. We classified accuracies in two terms. First, we have obtained accuracies of these each algorithm are compared them. Then we have used sklearn while designing the neural network

model. The sklearn model for Multi-layer perceptron has various solving and activation function. So we have also compared the accuracy of each solving technique combined with each activation function defined.

IV. THE COMPUTATION USED IN EACH MODEL

A. Single Layer Perceptron

output = w.x + b
it is 1 if w.x+b > 0
it is 0 if w.x+b < 0
b is threshold between 0 or 1
w is weignt between node
x is input to the node

B. Voted Perceptron

$$V = V + yX$$

X denotes the Euclidean length of x
y denotes the labels in {-1, 1}
V denotes the prediction vector
predicts the label of new instances:

$$\hat{y} = \text{sign}(V. X)$$

C. Multi-Layer Perceptron

The computation is divided in two parts:
Hidden Layer:

$$H = Wi . Xi + b$$

Wi – weight of the node ranges from 1 – i
Xi - the number of input to the node ranges from 1-i
b – the bias node weight
The Activation function

1) Sigmoid Function(logistic):

$$f(H) = 1 / (1 + e^{(-H)})$$

2) Tanh Function(tanh)

It gives the hyperbolic vlaue for the intput

$$f(x) = \tanh(x).$$

3) Rectified Linear Unit Function(relu):

It return max between two values

$$f(x) = \max(0, x)$$

4) Linear Bottelneck function:

it has no operation. So it is named as no-op activation funtion

$$f(x) = x$$

The different solving technique that are used in model:
lbfgs : The family of quasi Newton Methods
sgd: The stochastic gradient descent
adam : stochastic gradient based optimizer

V. CONCLUSION

The obtained accuracy of the model are as follows:

Name	Instances	Accuracy	Error
Single Layer Percepton	11055	93.05	6.95
Voted Perceptron	11055	93.24	6.739
Multi-layer perceptron	11055	97.45	2.46

The accuracies found different the possible reasons for reduction of accuracy are as follows:

1) Single Layer Perceptron

The accuracy is less beacuse the instances present on non-separable regions. It increased error in model prediction. The time taken to build the model is fast

2) Voted Perceptron

The voted perceptron accuracy is less because of the prediction vector. The time taken to build model is very fast as compared to other two techniques.

3) Multi- Layer Perceptron

The accuracy of this model is very high because of optimization of weight till the highest accuracy achieved. The fault tollerence of this model is very high. The backpropogation included in the model increase the accuracy. The time taken to build the model is based on the learning rate of the model. It can be increased with respect to the increase in the learning rate but this may reduce the accuracy of the model

We have further reaserched on the accuracy of this model using differnt activation function and solving technique defined in the sklearn.

Solver	logistic	Tanh	relu	identity
lbfgs	55	96	97	92
sgd	55	94	55	92
adam	93	96	96	92

We have also analysed the confusion matrix for each model.

REFERENCES

[1] Supervised fication Machine Learning: Techniques[2007] Review of classification technique
[2] Parametering detection optimization using for adaptive intellegient neuron phish-fuzzy[2014]
[3] Multi-Layer Perceptron
[4] A Comparative Study of Training Algorithms for Supervised Machine Learning[2012]:
[5] Deep learning made easy bylinear transformation in perceptron
[6] Scikit Learn: Machine learning in python[2011]: The Journal of machine learning volume 12
[7] PhishWho- Phishing Webpage Detection via identity keywords extraction and target domain name finder[2016]
[8] Phishing website detection based on supervised machine learning with wrapper feature selection[2017]:International Journal of Advanced Computer Science and Applications,Vol. 8, No.9
[9] Detection and prediction of phishing website using classification of mining teachnique[2016]:International Journal of Advanced Computer Science and Applications,Vol. 147, No.5
[10] Detecting phishing emails using text and data mining[2012]:ieeexplore
[11] Detecting redirection spam using multi-layer perceptron[2017]
[12] Efficient supervised prediction learning of phishing web algorithm[2012]
[13] Phishing website prediction using classification technique
[14] Evolving fuzzy neural network for phishing email detection[2017]
[15] Mulilayer Perceptron: Architecture optimization and training
[16] Behaviour analysis of multilayer perceptron with multiple hidden neurons and hidden[2011]
[17] Detection of phishing webpages using hetrogeneous transfer learning[2017]

[18] Phishing Web Website Scraping and Detection Data Framework Mining[2017]

