International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I3P106

Volume 69 Issue 3, 26-29, March, 2021

Original Article

Phishing Website Detection using Ensemble Technique

Priyanka Sharma[1], Rajni Ranjan Singh Makwana[2]

[1]*Madhav Institute of Technology & Science, Gwalior, M.P., India*

[2]*Madhav Institute of Technology & Science, Gwalior, M.P., India*

Received Date: 28 January 2021
Revised Date: 15 March 2021
Accepted Date: 17 March 2021

Abstract *- In the age of the internet, there is an enormous number of online transactions performed every day; therefore security and privacy of online transactions and banking websites is a challenging task. Website phishing attacks are carried out by presenting a fake website as a genuine one in order to gain confidential information and using that for some non-genuine activities. In this work, the Bagging technique is used with neural network and LMT classifiers as base classifiers in ensemble to classify a set of URLs and to determine the URLs as phishing or legitimate so that a user can be secured from phishing attacks. In this work, we have obtained an accuracy of 90%.*

Keywords *- Phishing detection, Ensemble classifiers, Classification techniques, Internet security, Machine Learning.*

I. INTRODUCTION

Sometimes even well-educated and trained users can also be trapped hence opening a phishing website that looks authentic or secure and giving up on sensitive information. Uniform Resource Locators (URLs) can be a general approach in terms of finding the malicious website. Through URL, a document can be addressed around the World Wide Web [1].

In this work, a neural network classifier has been used with voting techniques to obtain a system that can identify fresh URLs as legitimate or phishing ones. Neural networks with other techniques accept inputs, train the dataset, and the output layer contains the result, which shows if the URLs are legitimate or phishing.

This study also compares the proposed work with previous work, as the accuracy has been improved for the respective classifier.

II. RELATED WORK

Till now, many authors proposed different methods for the topic "phishing website classifications." These methods are based on different approaches. A few of them are described as below:

Patil and Patil's [2] survey shows a basic overview of detection techniques for malicious webpages covering various types of web-page attacks.

Hadi, Aburrub, and Alhawari [3] used the Fast-Associative Classification Algorithm (FACA) for detecting phishing URLs. By using FACA, all frequent rule item sets can be discovered, and a model for classification can be built. In which a data set has been explored with two classes, legitimate and phishing.

In addition, Arun Kulkarni1, Leonard L. Brown, III2 [4] proposed an approach using MATLAB scripts with four classifiers, which are the Naïve Bayes' classifier, decision tree, Support Vector Machine(SVM), and the Neural Network. To detect phishing URLs, these classifiers were implemented. Firstly features are extracted from the URLs, and then URLs have been classified via the model developed by training set data.

III. PROPOSED APPROACH FOR PHISHING WEBSITE CLASSIFICATION

In this work, an approach has been proposed using Ensemble. The flow diagram of the approach is as follows:

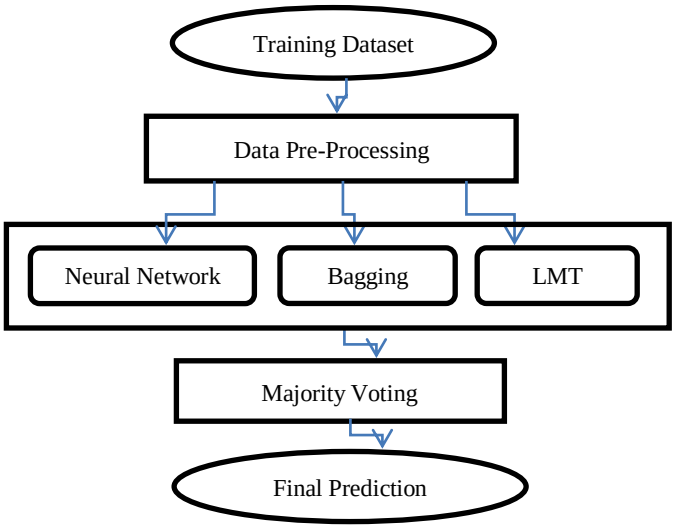

Fig. 1 Proposed Approach

A. Dataset

In this work, a phishing dataset from the University of California, Irvine Machine Learning Repository, and Center for Machine Learning and Intelligent Systems [5] has been utilized for testing and training purposes. It contains features from 1353 URLs, from which 548 URLs are legitimate, 103 are suspicious, and 702 are phishing. The features are described as follows:

a) ***Server Form Handler (SFH):*** Generally, where the webpage is loaded, the information process happens in the same domain. From the handler, the server is either transformed to another nonlegitimate domain or empty in phishing websites.

b) ***Secure Socket Layer (SSL):*** HTTPs protocol might not use by Phishing websites. So, it warns the end-user so that they know the particular site is secured by SSL or not.

c) ***Popup windows:*** Generally, popup windows are used to pop notifications and not to ask the users their credentials if the site is a legitimate one.

d) ***Request URL:*** Generally, the domain for loading the webpage is the same as the objects are loaded from legitimate websites.

e) ***URL of the anchor:*** The hypertext reference is used for the target definition of the anchor element. A website is defined as suspicious or phishing if that anchor element is not pointing to the domain where the webpage is loaded but a different domain.

f) ***Web traffic:*** High web traffic means the website has regular visitors and shows that it is legitimate.

g) ***URL length:*** Often, long URLs are used to hide the suspicious part of it phishing websites.

h) ***Age of the domain:*** Mostly, legitimate domains are the ones that are serviceable for a long time.

i) ***The URL having an IP address:*** Generally, URLs don't have an IP address, hence having one in the domain name depicts that the website can be suspicious.

j) ***Class:*** There are three classes in this data set: suspicious, phishing, and legitimate, in which the URLs are categorized.

B. Data Preprocessing

On the dataset, data preprocessing has been applied to make it a knowledgeable information set. After preprocessing next step is to balance the dataset through the class balancing technique, which is described below:

Class Balancing: To fix the imbalanced data and to trend the data equally, class balancing is used. To make the data balanced, oversampling and undersampling for instances can be done, of the minority class and the majority class, respectively.

C. Classifiers

In this work, a neural network classifier has been used along with the voting technique. Both are described as below:

Neural Network: For statistical classifiers, neural networks provide a great alternative. A training set of data is used by neural networks learning, and then they make decisions [6, 7]. A neural network has layers as units that take input and generate some output and pass it on to the consecutive layer.

LMT: LMT, known as a logistic model tree (LMT), is a model with an associated supervised training algorithm that is used for classification. The algorithm actually combines logistic regression (LR) and decision tree learning [8, 9].

D. Classification Methods

Ensemble: By means of ensemble technique, advantages of the different algorithm can be integrated, and through which one optimal result can be achieved [10]. Therefore, the ensemble is to combining multiple models to train the dataset in a way to improve the accuracy. In this paper, there are two methods utilized to the ensemble:

Voting: For combining classifiers together, a voting technique is used, which is a class that makes classifiers as one unit to produce best estimate results.

Bagging: Another method that has been implemented in this research is bagging, which is another method of the ensemble. Bagging is also called Bootstrap aggregating. It's designed to enhance the accuracy of machine learning algorithms. It also minimizes the variance, therefore

abolishing fitting. For the model averaging method, bagging is a special case.

IV. RESULTS

This section represents the preprocessing and classification process. An experiment has been carrying out using the Weka interface.

The data set that has been used for this work from The University of California, Irvine Machine Learning Repository has nine attributes, and it's 1,353 samples. The histogram shown in Fig. 2 depicts the values for the data set. In that, three peaks are visible, which are representing three classes and their counts, label 0, 1, and -1 representing suspicious, legitimate, and phishing URLs, respectively.

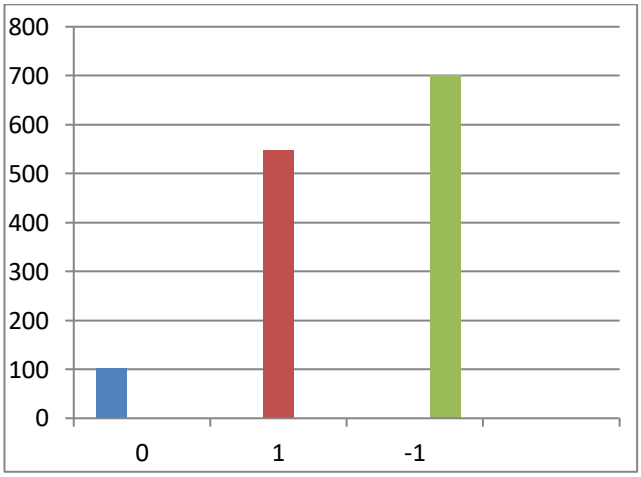

Fig. 2 Dataset histogram before class balancing

Fig. 3 represents a graph containing values after class balancing has been applied to the given data set.

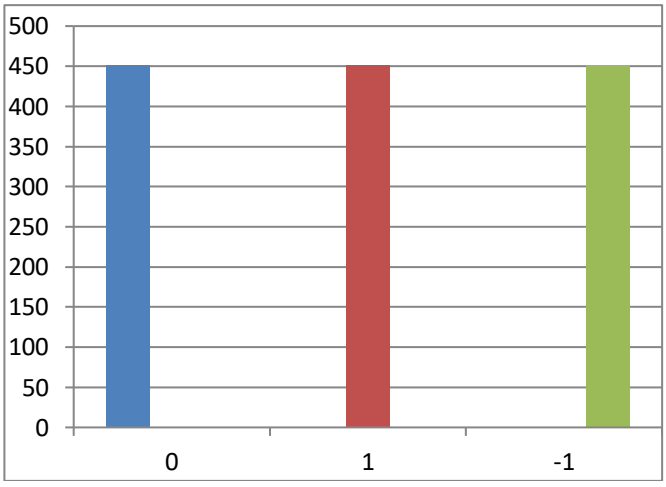

Fig. 3 Dataset histogram after class balancing

There are nine units within the neural network; one unit is assigned to every feature of the input layer. The hidden layer contains ten units, and the output layer consists of 3 units for three classes each, as mentioned within the above histogram.

In this work, further, some techniques are employed in order to improve the consistency of the neural network classifier and compared with the previous work's same classifier.

For which the dataset has processed through class balancing followed by classification through voting technique, within which the data processed through percentage split, where 60% and 40% of the samples are selected randomly for training and testing, respectively.

The results contain the accuracy of the classifiers, True Positive Rate (TPR), and False Positive Rate (FPR) for phishing URLs, as shown in Table 1.

During classification, different combinations and permutations have been applied and tested before preferring to the high accuracy combinations, as follows:

Table 1. Results for the tested Proposed Technique

Classifiers	TPR	FPR	Accuracy
Neural Network+Bagging+LMT	90.5%	5%	90.49%

Also, if compared to the previous work, current work has been acquired the higher consistency for the same classifier, as shown in Table 2.

Table 2. Comparison in between the previous and proposed work

Method Proposed by Kulkarni & Brown[4]		Proposed Methodology	
Classifiers	**Accuracy**	**Classifiers**	**Accuracy**
Neural Network	84.87%	Neural Network+Bagging+LMT	90.49%

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In this work, the Neural Network and LMT have been used as a base classifier in ensemble to make a classification model with class balancing, bagging, and voting techniques. This classifier model can be used to detect phishing URLs, hence helping a user to protect the system from web phishing attacks. Two steps are evolved in phishing URLs detection, during which extracting features from the URLs and URL classification using the training set data developed model are included. The data set has been utilized in this work, provided the extracted features.

This work shows higher accuracy of the combined classifiers, which is 90.49% which has been achieved by applying the

percent split feature. Additionally, by adding more classifiers in this combination, together with different split percentages, the accuracy will be higher.

B. Future Work

Using a few more and different combinations for the classifiers within the voting technique can lead to improvements in accuracy values for this classifier. Also, using the frequent item data sets with the minimum support and confidence values, associative rules will be generated; hence to classify URLs using associative rules, it builds a rule-based system. Then this may be compared with other classification methods.

Another approach is to part the feature space using fuzzy membership functions then bring out and enhance classification rules to generate the classification rules from the samples of data [12]. These extracted rules are often utilized to make a fuzzy inference system to classify URLs.

REFERENCES

[1] N. Lord, What is a Phishing Attack? Defining and Identifying Different Types of Phishing Attacks. https://digitalguardian.com/blog/what-phishing-attack-defining-and-identifying-different-types-phishing-attacks, (2018).

[2] D. R. Patil and J. Patil, J., Survey on malicious web pages detection techniques, International Journal of u-and e-Service, Science and Technology, 8(5)(2015) 195–206.

[3] W. Hadi, F. Aburrub, and S, Alhawari, A new fast associative classification algorithm for detecting phishing websites, Applied Soft Computing 48(2016) 729-734.

[4] Arun Kulkarni1, Leonard L. Brown, III2, Phishing Websites Detection using Machine Learning, (IJACSA) International Journal of Advanced Computer Science and Applications,10(7)(2019).

[5] UCI Machine Learning Repository: Website Phishing Data Set (Online) https://archive.ics.uci.edu/ml/datasets/Website+Phishing.

[6] R. P. Lippman, An introduction to computing with neural nets. IEEE ASSP Magazine,3(4)(1987) 4-22.

[7] S. L. Gallant, Neural network learning and expert systems, The MIT Press, Cambridge, MA, (1993).

[8] Niels Landwehr, Mark Hall, and Eibe Frank., Logistic model trees,(2003).

[9] Landwehr, N.; Hall, M.; Frank, E., Logistic Model Trees. Machine Learning., (2005).

[10] Xianwei Gao, Chun Shan, Changzhen Hu, Zequn Niu, Zhen Liu An Adaptive Ensemble Machine Learning Model for Intrusion Detection. IEEE,7(2019). https://en.wikipedia.org/

[11] A.D. Kulkarni, Generating classification rules from training samples, International Journal of Advanced Computer Science Applications, 9(6)1-6.

[12] Surendiran,R., and Alagarsamy,K., Privacy Conserved Access Control Enforcement in MCC Network with Multilayer Encryption. SSRG International Journal of Engineering Trends and Technology (IJETT) 4(5) (2013) 2217-2224.