2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT 2020)

Gwalior, India
10 – 12 April 2020

IEEE Catalog Number: CFP2018P-POD
ISBN: 978-1-7281-4977-6

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP2018P-POD
ISBN (Print-On-Demand):	978-1-7281-4977-6
ISBN (Online):	978-1-7281-4976-9
ISSN:	2329-7182

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

2020 9th IEEE International Conference on

Communication Systems and Network Technologies

CSNT 2020

Table of Contents

Track-A: Microwave Components, Antenna and Propagation

Track-B: AI, Fuzzy and Machine Learning

Track C: Mobile Computing, WSN, IOT and next Generation Networks

Track D: Hardware Design, Data Mining and systems

Track E: Cryptography, Blockchain and Security Algorithms

Track F: Image Processing, Signal Processing intelligent systems

Phishing Detection Using Significant Feature Selection

D.N. Goswami
S.O.S.in Computer Science And Applications, Jiwaji University
Gwalior, India
e-mail address:
goswamidn@yahoo.com

Manali Shukla
S.O.S.in Computer Science And Applications Jiwaji University,
Gwalior , India
e-mail address:
shukla_manali@rediffmail.com

Anshu Chaturvedi
Department Of CSE & IT, Programme M.C.A
Madhav Institute Of Technology &Science
Gwalior, India
e-mail address:
anshu_chaturvedi@yahoo.co.in

Abstract— ***Growth of cyber attacks is rapidly increasing in the entire world. To provide prevention from these attacks is a great challenge for the experts. Intruders are keep on adapting new methods and techniques to carry out their malicious goals. Phishing plays a dominant role in the field of web attacks and it has been used as a weapon by the attackers. In this paper we have given two algorithmic approaches to the problem of Phishing identification with reduced number of attributes. It makes this approach simple yet efficient. The first algorithm assigns weight to all attributes with respect to uniform resource locators. We have employed various analysis mechanism to identify significant role of selected attributes for the purpose of Phishing identification . The second approach takes former's output as input and classifies the uniform resource locators labeling as phishing or non phishing. The experimental work verifies that the approach for phishing detection proposed in this paper can attain a high accuracy in comparison to existing algorithms.***

Keywords—*cybercrime; Phishing; Phishers; subdomain ;* **url;** *Weka;*

I. INTRODUCTION

We cannot deny the existence of phishing all over the globe. It becomes an easy and effective tool for the intruders to perform cyber scams. Phishing has been implemented using several digital channels such as email, message based services, social platforms and websites. Millions of Phishing attack has been reported by several sources. The Anti-Phishing Working Group (APWG) in [1] identified presence of total 182,465 phishing sites in the second quarter of 2019 which was observed 180,768, during the first quarter of 2019 but there is a huge increment being observed in phishing site's growth from 138,328 in fourth quarter of 2018 and the 151,014 in the third quarter of 2018. Therefore existence of Phishing websites becomes an important concern among the web security domain experts.

Various approaches have been proposed in the past decades in order to provide prevention from phishing websites. Phishing website identification can be carried out using content and non content based techniques. The authors in [2] defined the meaning of content and non content based approaches. In content based approaches Phishing is identified by analyzing contents of website which includes features such as links, spelling errors, passwords etc on the other hand non – content based approaches are mainly relies on uniform resource locators features and host based features such as length of uniform resource locators and presence of some special characters , presence of internet address etc.[2]. Blacklist and White list based approaches were also practiced to determine the presence of phishing websites. Blacklist contains list of websites which already reported as phishing websites and white list contains websites not being reported as Phishing [3]. Several Researchers found that these approaches suffers from the limitation of on time updating of lists as well as cannot fight against zero day phishing attacks

Phishing can be defined as web attack which is implemented by the intruders to gain access of personal data of users over the internet which helps them to perform financial frauds and scams. Therefore, Identification of Phishing websites becomes an emergent field of research among various researchers. We have made following contributions in this research paper (i) Illustrating the impact of using significant Uniform resource locators based features in the process of phishing website classification (ii) Examine the results obtained after using uniform weight assignment to all the features being selected. (iii) Evaluated the performance of the proposed approach with the existing algorithms (iv) Summarized the results based on performance evaluation measures The next section describes literature relevant to the problem of phishing identification along with the pros and cons of algorithms used in the comparative analysis of the proposed work

II. RELATED WORK

In this section we analyzed and reviewed some of the research contribution made by the researchers in the past. Many researchers paid attention on the fact that cyber attacks such as Phishing employs mechanism that utilizes factors related to

978-1-7281-4976-9/20/$31.00 ©2020 IEEE
DOI: 10.1109/CSNT.2020.55

user ‘s limitations .They also identified that attackers aim all the vulnerable aspect which are present due to the weaknesses shown by web users[4]. The author in [5] proposed an approach which is based on term frequency .CANTINA determines whether a web page is legitimate or illegitimate. The author in [6] proposed a Bayesian classification system for the identification of suspicious URL. Bayesian classification is easy to implement but in some cases it suffers from the limitation of conditional independence of class which may results into loss of accuracy. The author in [7] determine presence of illegitimate websites using WHOIS database. The website being identified as malicious is removed from the host server through sending notification. To determine the presence of malicious uniform resource locator author in [8] utilizes features based on structural and lexical aspect which are found in the URL’s by employing SVM based approach. It is observed that SVM may requires various key parameters which are needed to be initialize correctly in order to gain better classification results. The author in [9] identified forty two features out of one hundred seventy seven using Correlation Features Set, Wrapper methods and employed different algorithms. In Our opinion this approach used large number of features which would increase execution and processing time of algorithm. It is always better to have reduced number of attributes. A study on Feature relevant to the websites have been done by the authors in [10] and they organize features in the form of clusters. Their study may generate biased result due to imbalanced use of dataset. In another study in [11] authors employed characteristic features of legitimate websites in order to develop a model based on rules for determining phishing attack. To develop a new method for detecting phishing attacks authors in [12] explored data mining techniques. Authors in [13] proposed a framework for the identification of malicious website and extracted features of uniform resource locators using subset based selection technique. They also employed machine learning based algorithms. In an another study [14]. Authors have used reduced features set and suggested “Fresh-Phish which is an open-source framework for the identification of phishing website Authors in [15] suggested feature based mechanism in order to classify uniform resource locators as phishing and non phishing. Feature used in this approach includes lexical features, Alexa rank, Page Rank , WHOIS features and Phish tank features. Although there is little concern needs to be taken when ranking features are used because the websites with relatively low traffic measures will not be correctly ranked by Alexa which may lead to inaccuracy. Authors in [16] defined ZeroR algorithm which mainly relies on the use of frequency based approach. Frequency table which is based on targets and neglects rest of the predictors. In other words we can say that it majorly makes predictions for the majority class. This algorithm suffers from the limitation of less predictably. OneR classification algorithm is more effective than the ZeroR algorithm and generates one rule for each attribute It handles missing and numeric attributes. Generates better and more accurate rules of classification than ZeroR. In this approach Nominal attributes suffers from the problem of overfitting and it selects attribute randomly when error rate is equal[17]. In [18] authors have identified attributes which plays significant role in identifying maliciousness of website. Next section will describes the methodology employed in the proposed research work.

III. PROPOSED WORK

In the proposed approach, we have used attributes based on uniform resource locators in order to identify presence of phishing website. Selection of these features is based on extensive analysis performed to determine the significance of selected features of several Uniform Resource Locators. One such analysis results are discussed later in the result section which is implemented in WEKA using dataset of UCI machine learning repository [19]. This dataset contains 11055 instances out of which existence of few features which we have selected is being examined. We have examined several other uniform resource locators data sets which we prepared using Phishtank[20], it is used as an anti – Phishing website which provides services to verify suspicious Phishing sites. Extraction of selected features is implemented using python IDE. We have also used some attribute selection filter such as CfsSubsetEval which determines worth of attributes in a subset of attributes on the basis of self predictive ability of individual attribute. In the proposed research work, our objective is to attain satisfactory accuracy with optimum number of attributes in order to identify Phishing websites. In this section, we are presenting an algorithm for assignment of weight for each attribute which plays significant role in turning a benign URL into a malicious URL which we have identified after performing deep analysis and testing on the attributes responsible for such type of attacks.

The proposed WA_algorithm takes featuresSet as input which contains values of all the extracted attributes from the dataset of uniform resource locators. In this approach, weights are assigned uniformly and our approach is to give emphasis on each feature of uniform resource locators which furthers serves as input to the classifier. Pseudo code of the proposed WA_algorithm is given in this section which assigns the weight to each URL by assessing the presence of each feature which appears to be malicious. The WA_algorithm increments the weight of the corresponding URL each time if the conditional statement becomes true and it decrements the weight of the corresponding URL only if none of the conditions appear to be true and it assigns two values for the weight of URL i.e. positive and negative where positive weight signifies the presence of maliciousness of uniform resource locators and negative

weight identifies the absence of maliciousness of uniform resource locators.

In our classification approach, we are considering only those URLS as benign which are having negative value of weight and URL's with positive values will be considered as malicious. The WA_algorithm outputs featureSet with the weight attribute which is used as input in our classification approach of URL's .Our classification approach works on the output provided by our proposed WA_algorithm as input dataset containing the values of all the features extracted from the URLs along with their respective weight assigned by WA_algorithm. In our classification approach, We have classified the URL's into phishing and non – phishing labels after assessing this weight parameter of the corresponding URL. Our classification methodology evaluates this weight parameter to determine whether it has positive or negative weight which signifies either the presence or absence of certain malicious features of URL's respectively.

This classification approach works simply yet provides efficient results in comparison to existing algorithms using evaluation criteria's e.g. accuracy, TPR etc which is discussed in the results section of this paper. Pseudo code of proposed weight assignment algorithm is given in this section in the form of IF- Then rules in figure no.1.

```
WA_algorithm(Weight Assignment Algorithm)
Input: FeatureSet
Output: FeatureSet with weight
Set weight = 0

IF(no of dots > SetVal)  then
Weight = Weight + 1
EndIF
IF(Presence of hyphen >= SetVal) then
Weight = Weight + 1
EndIF
IF(len of url >= SetVal) then
Weight = Weight + 1
EndIF
IF(presence of at = = SetVal) then
Weight = Weight + 1
EndIF
IF(presence of double slash= = SetVal) then
Weight = Weight +1
EndIF
IF(no of subdir >=SetVal) then
Weight = Weight + 1
EndIF
IF(no of subdomain > = SetVal ) then
Weight = Weight + 1
EndIF
IF(len of domain > = SetVal ) then
Weight = Weight + 1
EndIF
IF (is IP  = =  SetVal ) then
Weight = Weight + 1
EndIF
IF( Weight = = 0) then
Weight = Weight – 1
EndIF
Exit.
 Else
                weight = weight – 1
 Exit
                                END
```

Figure 1: Pseudo code Of proposed Weight Assignment Algorithm

IV. RESULTS

In this section, results are analyzed and compared with existing algorithms This section describes the evaluation of the proposed work which has been done using dataset containing 1500 URL's including both malicious and non malicious URL's which we have collected from PHISHTANK repository [20]. The dataset is split into 70-30 ratio used for training and testing the proposed classifier accuracy respectively. The performance of the proposed classifier is evaluated using different measures which are computed as given in the respective tables and these measures are discussed below –

(i) True Positive: The total number of Uniform resource locators correctly classified as malicious or phishing.

(ii) True Negative: It is the total number of correctly classified uniform resource locators as legitimate.

(iii) False Negative: It is the total number of malicious uniform resource locators incorrectly classified as legitimate.

(iv) False Positive: It is total number of legitimate or genuine uniform resource locators incorrectly classified as phishing which is also known as false alarm rate . It should be reduced in order to achieve accuracy.

(v) Accuracy : Accuracy = (True Positive + True Negative) /(True Positive + True Negative + False Positive + False Negative).

(vi) Rate Of Error : (False Positive + False Negative)/ Total

We have identified the significance of some selected features using Visualization of attributes in WEKA Snapshot given in figure no. 2 and figure no.3 clearly illustrates the significant role of uniform resource locators based attributes in the Phishing website identification.

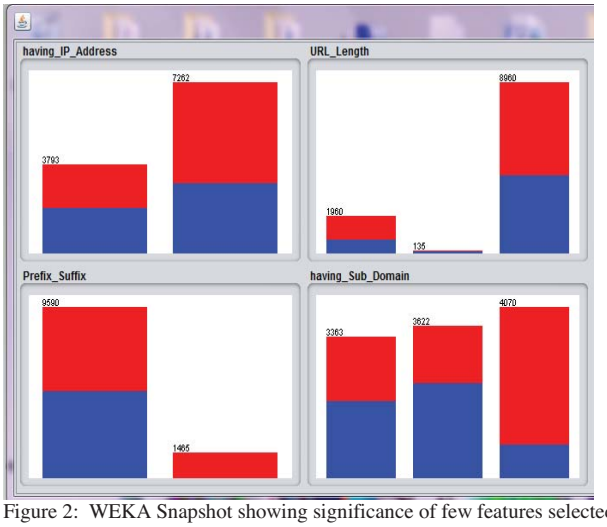

Figure 2: WEKA Snapshot showing significance of few features selected in the proposed work.

In the Figure 3 Internet protocol address attribute with (-1 1) values is shown along with the class attribute result having
(-1 1) values indicating phishing and legitimate sites respectively. If internet protocol address value is illegitimate(-1) then there will be more chances of phishing which is indicated using blue area and if this feature is found legitimate(1) there will be more chances of non phishing which is indicated by red area in figure no 2.

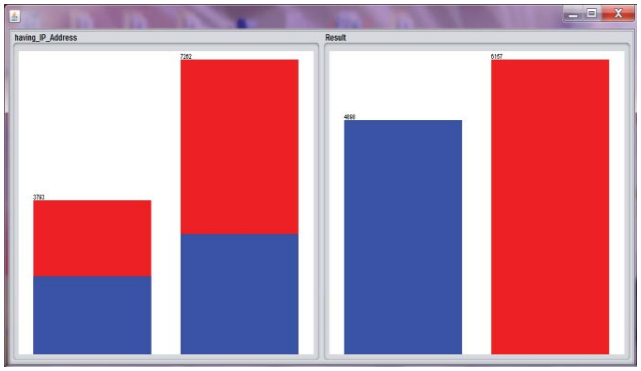

Figure 3 : WEKA Snapshot showing significance of Internet Protocol address attribute

Table no. 1 reflects satisfactory performance of our proposed classification approach on the basis of performance measures- rate of True positive, True negative , False positive , False negative and error rate.
We have compared rate of true positive and False Positive with the ZeroR and OneR algorithm which is depicted in Table no. 2 which shows that our proposed approach is giving better rate of true positive and reduced rate of false positives in comparison of ZeroR and OneR algorithms.
The results shown in Table no. 2 clearly reflects that the proposed approach performs comparatively better on the basis of specified measures and these results are also represented using a chart in figure no 4

Algorithms	Accuracy
One R	49.3%
Zero R	51%
Proposed Approach	74.4%

Table 1: Calculated values of Rate of True Positive , True Negative , False Positive , False Negative and Error Rate of Proposed approach

Table 2: Comparison of calculated Rate of True Positive and False Positive of proposed Approach with zeroR and OneR algorithms.

Existing Algorithm	TPR	FPR
ZeroR	0.511	0.511
One R	0.493	0.485
Proposed Approach	0.76	0.272

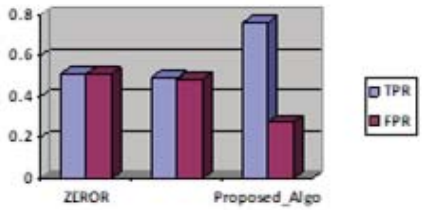

Figure 4: Comparision of True positive rate and False Positive

Fig. 4 shows Comparison Chart of True Positive rate and False Positive rate of proposed Approach with the ZeroR and OneR Algorithms. Table no. 3 reflects that the proposed algorithm performs better than the ZeroR and OneR algorithms on basis of accuracy measure which comes out 74.4% which is greater than that of existing algorithms zeroR (51%) and OneR (49.3%) . Relative chart is given in figure no.5. which represents comparison of proposed approach with the ZeroR and OneR algorithms using accuracy measure.

Table 3. Comparison of proposed Approach using Accuracy measure with ZeroR and OneR algorithms.

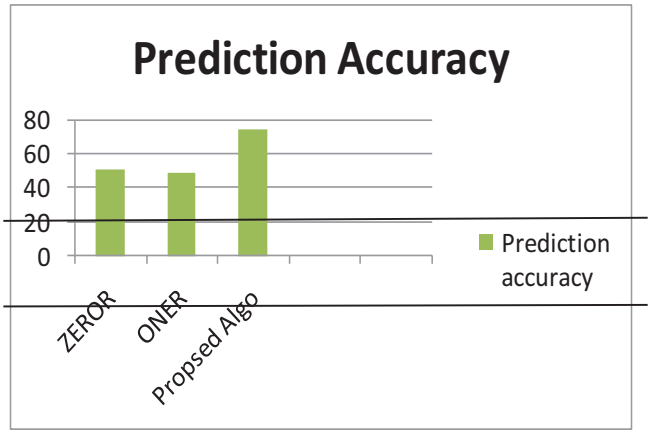

Fig. 5: Comparison of Proposed approach with the ZeroR and OneR Algorithms.

.

Analysis of all these performance measures illustrates the fact that the methodology proposed in this paper with optimum number of attributes can give better performance. Chart comparing accuracy measure of Proposed approach with the ZeroR and OneR Algorithms has been shown in Figure 5, which shows the clear supremacy of the proposed work..

V. CONCLUSION

This research paper illustrates comparative analysis of the proposed work with some existing algorithm on the basis of standard performance measures. Proposed approach has shown significant performance with optimum number of attributes in comparison of existing algorithms. False alarm rate is always been an concerning issue for which our proposed approach of uniform weight assignment and classification has shown great reduction in false alarm rate in comparison of some existing algorithms.

REFERENCES

[1.] APWG, et al., second quarter 2019. Phishing activity trends report. Anti- Phishing Working Group. URL https://docs.apwg.org/reports/apwg_trends_report_q2_2019.pdf.

[2.] Afroz, S.; Greenstadt, R., "PhishZoo: Detecting Phishing Websites by Looking at Them", IEEE Fifth International Conference on Semantic Computing (ICSC), 2011, pp. 368-375

[3.] Weiping Wang, Feng Zhang, Xi Luo and Shigeng Zhang "PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural networks" Hindawi Security and Communication Networks Volume 2019, Article ID 2595794, 15 pages.

[4.] Mahmoud Khonji, Youssef Iraqi and Andre Jones, "Phishing Detection: A Literature Survey", IEEE Communication Surveys & Tutorials, vol. 15, No. 4, pp.2091-2121, 2013.

[5.] Yue Zhang, Jason Hong, Lorrie Cranor, "CANTINA: a content-based approach to detecting phishing web sites", in Proceedings of the International World Wide Web Conference, pp.1-10, 2007.

[6.] Chia-Mei Chen, D. J. Guan, Qun-Kai Su, "Feature set identification for detecting suspicious URLs using Bayesian classification in social networks", Elsevier Information Sciences 289 (2014) 133–147

[7.] Shah R, Trevathan J, Read W, Ghodosi H (2009): A proactive approach to preventing phishing attacks using Pshark. In Sixth international conference on information technology new generations. IEEE, Las Vegas, pp 915–921

[8.] Huang H, Qian L, Wang Y (2012) :A SVM based technique to Detect phishing URLs. Int Technol J 11(7):921–925.

[9.] R. Basnet, A. Sung, and Q. Liu, "Feature selection for improved phishing detection," Advanced Research in Applied Artificial Intelligence, pp. 252–261, 2012.

[10.] Mohammad R., Thabtah F, McCluskey L (2012) An Assessment of Features Related to Phishing Websites using an Automated Technique. 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012); 2012; London: ICITST.

[11.] Ramesh G, Krishnamurthi I, Kumar K. An efficacious method for detecting phishing web pages through target domain identification. Decision Support System 2014; 61: 12-22.

[12.] Ajlouni M, Hadi W, Alwedyan J. "Detecting phishing websites using associative classification" Journal of Information Engineering and Applications 2013; 5: 1899-1905.

[13.] Mustafa AYDIN, Nazife BAYKAL : Feature Extraction and Classification Phishing Websites Based on URL : IEEE,2015

[14.] Hossein Shirazi, Kyle Haefner, Indrakshi Ray: Fresh-Phish: A Framework for Auto-Detection of Phishing Websites: In (International Conference on Information Reuse and Integration (IRI)) IEEE,2017

[15.] Bhagyashree E. Sananse, Tanuja K. Sarode : Phishing URL Detection: A Machine Learning and Web Mining-based Approach: In International Journal of Computer Applications, 2015

[16.] Andreeva, P., M. Dimitrova, P. a Radeva (2004), "Data Mining Learning Models and Algorithms for Medical Applications", 18 Conference Systems for Automation of Engineering and Research (SEAR), Bulgaria

[17.] Buddhinath, G., D. Derry (2003), "A Simple Enhancement to One Rule Classification", Department of Computer Science & Software Engineering University of Melbourne.

[18.] Abdelhamid N and Ayesh A: Phishing detection based associative classification data mining Expert Systems with Applications 41(13) pages 5948- 5959, Oct 2014

[19.] Lichman, M. (2013). :UCI Machine Learning Repository University of California, School of Information and Computer Science. [http://archive.ics.uci.edu/ml].

[20.] Data for Fraud websites, https://www.phishtank