Journal of Science and Technology
ISSN: 2456-5660 Volume 5, Issue 6, Nov-December 2020

www.jst.org.in *DOI: https://doi.org/10.46243/jst.2020.v5.i6.pp11- 18*

A Survey on Phishing Detection and The Importance of Feature Selection In Data Mining Classification Algorithms

Shikha Verma[1], Arun Kumar Gautam[2]
[1]*(Computer Science, Ram Lal Anand College/ Delhi University, India)*
[2]*(Computer Science, School of Computer and System Sciences / Jawaharlal Nehru University, India)*
[1]*Corresponding Author: shikhaverma@rla.du.ac.in*

To Cite this Article
Shikha Verma and Arun Kumar Gautam "A Survey on Phishing detection and the importance of feature selection in Data Mining Classification algorithms", Journal of Science and Technology, Vol. 05, Issue 06, Nov-December 2020, pp11-18

Article Info
Received:18-06-2020 Revised:15-08-2020 Accepted:30-08-2020 Published: 22-09-2020

Abstract: *In this era of Internet, the issue of security of information is at its peak. One of the main threats in this cyber world is phishing attacks which is an email or website fraud method that targets the genuine webpage or an email and hacks it without the consent of the end user. There are various techniques which help to classify whether the website or an email is legitimate or fake. The major contributors in the process of detection of these phishing frauds include the classification algorithms, feature selection techniques or dataset preparation methods and the feature extraction that plays an important role in detection as well as in prevention of these attacks. This Survey Paper studies the effect of all these contributors and the approaches that are applied in the study conducted on the recent papers. Some of the classification algorithms that are implemented includes Decision tree, Random Forest , Support Vector Machines, Logistic Regression , Lazy K Star, Naive Bayes and J48 etc.*

Keywords*: Classification; Phishing attacks; Decision Tree; Random Forest ; Support Vector Machines; Logistic Regression ; Lazy K Star; Naive Bayes*

I. Introduction

Phishing is an online theft which aims to capture confidential information from the end user. It is the type of Scam in which the attacker uses bogus email or website to obtain and enter into the user confidential area. According to APWG, Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumers 'personal identity data and financial account credentials. Phishing can be implemented in different ways such as follows (Alnajim and Munro, 2009)

1. Email to Email - When the attacker uses email as the mode of gaining confidential information.
2. Email to Website - When the receiver gets the email embedded with bogus website.
3. Website to Website -When the end user clicks the phishing website via search engine.
4. Browser to website: - When someone misspelled a legitimate website and end up in entering phishing site which is semantically similar.

Phishing causes billions of dollars in damage every year and poses a serious threat to the Internet economy.[1] According to the Phishing Activity Trends Report, 1st Quarter 2019[2]

Table no 1: Statistical Highlights for 1st Quarter 2019

Count	Jan	Feb	March
Unique Phishing sites detected	48,663	50,983	81,122
Phishing email reports	34,630	35,364	42,399

According to the Latest study conducted by APWG.org Phishing Site and Phishing E-mail
Trends of the 1st Quarter 2019 were as follows:-

The total number of phishing sites detected by APWG in 1Q was 180,768. That was up notably from the 138,328 seen in 4Q 2018, and from the 151,014 seen in 3Q 2018.

Most studies target on increasing the accuracy of the website or email phishing detection by using different methods or techniques. There are several data mining algorithms that train and test the dataset such as Decision Tree, J48 , Support Vector machines, Naive Bayesian, Neural Network, Random forest, Logistic regression etc. Few researchers have also focused on the dataset features or the feature selection algorithms that helps in contributing to the accuracy of the phishing detection techniques. Results showed that Random Forest has outperformed best among the classification methods by achieving the highest accuracy 97.36% [3].Most researchers have used phishing dataset from UCI repository. The rest of the paper is structured as follows Section II talks about different finding from the recently conducted researches. Section III includes the Results and Discussion and Section IV concludes with the conclusion and future work.

II. Findings from the recent researches

Performance Comparison of classifiers on Reduced Phishing Website [4] .In this paper phishing website Dataset is taken from UCI machine learning repository which consists of 11055 records with 31 features [5].The following were the aim of this study:-
1. To reduce the dimension of the dataset for faster classification.
2. To show the performances of the classification algorithms.
The reduced dataset was classified into two parts. The first part was the training set and the second part was the testing dataset where 5-fold cross validation was applied. The algorithm used for the reduction of the dataset were:- Individual FS with 27 features , Forward feature selection with 24 features backward feature selection with 25 features plus I takeaway FS with 27 features and Association rule with 26 features. Comparative analysis was done where some of the feature selection algorithms were applied along with the implementation of the classification algorithms like Naive Bayes, Lazy K star, LMT, ID3, SGD, Bayes Net etc. The highest accuracy was produced by the Lazy K star classification algorithm with 97.58% accuracy on the reduced dataset with 27 features. Detecting Phishing websites using Data Mining [6] this paper proposed the solution of detecting phishing websites using data mining techniques. Few anti-phishing techniques includes list based, visual similarity based, content based and heuristic based techniques[7][8].List based keeps both the blacklist and the white list to check the URL entered after which it checks and sends the alert message to the user. In Layout similarity approach three metrics were involved for the check
1. Layout
2. Block Level and
3. Overall similarity check

In content based approach, examination of the content of a webpage was done to decide if it is genuine or fake. In heuristic based approach some of the common features like URL, HTML, and DOM etc were examined. The proposed system was a cloud based model which directly interacts with the chrome extension and was integrated with all the features of both client and server side. Various attributes were extracted from the URL fetched from the chrome extension and the classifier were used to predict whether the entered URL was malicious or safe.UCI repository dataset was used which had 11055 records out of which 4898 were phishing websites and 6157 were legitimate website. Features in the dataset were categorized as:-

1. Address bar based features
2. Abnormal based features
3. HTML and JavaScript based
4. Domain based features.

On Comparison with the classifiers used, the following results were obtained:-

Table no 02: Results

Method	Accuracy
Naive Bayes	92.980%
J48	95.752%
SVM	93.8037%
Neural network	96.9064%
Random forest	97.2592%
IBK Lazy Classifier	97.1777%

Final Results obtained showed that Random forest simplified classifier was found to have the highest accuracy.

Data mining a way to solve phishing attacks [9].There are many anti-phishing mechanisms to identify the phishing sites which prevent users from getting deceived. Some of the anti-phishing techniques include AntiPhish [10] and Webwallet [11] that helps the end-user from falling into the fraud or phishing attacks and some password hashing techniques like PwdHash, Password multiplier [12] and passpet [13].The proposed algorithm takes the raw Emails via Java application and extracts the text of the body and the message after which the text mining technique was applied. In the next stage clustering was done on similar feature which goes to the classifier algorithm that implemented the Naive Bayed classifier. The false positive rates of Gemini in different browsers were as follows:-

Table no 03: False positive rate of Gemini

Category	IE	Firefox	Chrome	Overall
Login	0.68%	0.68%	0.68%	0.68%
Phish	3.23%	0.39%	0.32%	0.64%

The false negative rates of Gemini were as follows.

Table no 04: False negative rate of Gemini

Category	IE	Firefox	Chrome	Overall
Login	0%	1%	0%	0.33%
Phish	0%	0%	0%	0%

Gemini also resulted in a very small overhead in loading the page and involved minimum overhead on browser

A Comparative Analysis and awareness survey of phishing detection tools [14].In this paper comparison of eight phishing detection tools are done to find out the best one. Some of the Phishing attacks are Email based phishing attack in which the email has the link attached with it which upon clicking redirects to a bogus website. Next is the Exploit based phishing attack in which the attacker performs the man-in-the-middle attack by changing the proxy settings.

The eight tools that are used for comparison are:-[15]

1. Phish detection
2. Netcraft
3. URL check info
4. Bit defender traffic light
5. Link extend
6. Anti Phishing
7. Spoof guard
8. Safe preview.

All these tools are implemented on the dataset of 2000 phishing websites and 500 legitimate websites and the comma parameters for the measurements are True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) and the accuracy. Accuracy was calculated for these eight tools and it was found that Anti phishing detection tool works out to be the best phishing detection tool with 94.32 % of accuracy followed by Bit defender Traffic light i.e. 93.92% and URL check info stands on third position with 85.72% accuracy.

Fine-Grained Mining and classification of malicious webpages [16].In this paper the author has built a classifier to extract variety of webpage features and has used machine learning algorithm to identify all possible threat types. The feature are extracted from the HTML contents, associated JavaScript code and corresponding URL. Dataset includes 1000 benign webpages and 1500 malicious webpages for experiment. The method of malicious webpage detection was classified into three categories:-

1. Honey Client [17]
2. Static detection
3. Machine learning based method

In the dataset selection four types of webpages were collected normal, phishing, spamming and malware. The dataset for phishing webpages was obtained from Phish Tank similarly the dataset of spam webpage was collected from WEBSPAM-UK 2007, for malware from malicious website labs and for the normal webpages it was obtained from Yahoo. In total 30 discriminative features in three aspects namely (HTML, JavaScript and URL) are extracted. The feature extraction is implemented in C++.In the model designed in this paper firstly the features were extracted and filtered then they were normalized to avoid imprecision of classification and finally the classifier performance was evaluated on the basis of accuracy, True Positive(TP) and False Positive(FP).In this experiment two classification models were implemented the first one which was used to determine the type of the webpage and the other model to check whether it is malicious or not. Binary classifier was used to detect the whether the page is malicious or normal which on further applying multiple classifier helps to identify whether the malicious page is from the category of phishing or spam or malware.

Detection of phishing website using C4.5 Data mining algorithm [18].In this paper analysis of C4.5 (J48) data mining algorithm was implemented through WEKA tool. The two Anti-phishing techniques which were used are [19].

1. Non-Technical approach which involves legal solution and training people and
2. Technical Approach where blacklist and heuristic based methods were applied [20].

The model can be explained as follows: - for the analysis of C4.5 algorithm firstly collection of phishing websites were made from which features were extracted and training and testing dataset were created. Training of J48 algorithm was done on training set using decision tree classifier model and lastly prediction was made on testing data and it was evaluated on the basis of different parameters. There are two steps in the construction of classification model. [21]

1. Learning phase
2. Prediction phase

In order to prepare the training dataset information gain technique was implemented which helped in the formation of the root node having the highest Information gain and eventually leading to the formation of decision tree where the branches of the tree stored the value of the attribute and the leaves predicted the classes. There were around 300 websites for testing among them 200 were phishing websites 154 which were predicted as phishing and amongst 100 legitimate 94 were detected. The success rate was obtained to be 0.826 with an error of 0.173 and the accuracy which was trained with 750 instances was found to be 82.6%.Different systems in heuristic based approach such as PhishZoo [22], PhishNet and LinkGuard [19] were proposed to detect phishing websites.
Phishing websites classification using Association classification (PWCAC) [23].In this research new algorithm was created named PWCAC Phishing website classification using association classification for the detection of phishing websites. This algorithm works through 3 phases namely:-

1. Extracting frequent item sets
2. Creation of classification rules and
3. Predicting unseen examples

In the first phase it aims to build a classifier having countable class value, it includes fast vertical mining technique known as difference sets [24] which calculates difference between the item sets n and n-1 .In the second phase it generate rules one by one on the training dataset and if that matches with at least one training dataset it is added to the model. Any training instance matches the rule body and if its class value is deleted the rule is added to the model else the rule is removed. The confidence value was 50% and tenfold validation was used in the experiment .It was conducted on CPU Pentium Intel Core TM i5 2430M @ 2.40 GHz Ram 4.00GB on windows 7 64bit Operating system. The dataset was collected from UCI [25] repository which had 2466 instance and this dataset had fresh and recent data.

Table no 05: Description of Dataset

No of Instance	No of attributes	Class Legitimate	Phishing
2466	30	54%	46%

One of the strongest point on PWCAC was its capability to find multiple labels per rule. In the end the proposed work resulted in 0.46% F1-Score which increased from MAC model and the second best result obtained from MCAR yet the PWCAC model increased it by 0.81%.

On feature selection for prediction of phishing websites [26].This study has implemented (KMO) Kaiser Meyer -Olken test for feature selection and has tested the same on UCI phishing website dataset. Further LR and SVM method were used for validation of feature selection. There are 3 basic approaches [27] to detect phishing websites namely

1. Content based
2. Heuristic based and
3. Blacklist based

The dataset [28] that is used here consist of 11055 instances with 30 attributes or features. The algorithm starts by reducing features using Kaiser-Mayer-Olkin (KMO) where partial correlation metric is calculated and then KMO index per variable was used to detect which all features are not related and lastly overall KMO was calculated which helped to remove features if KMO index was less than

0.751.Thus resulting into a total of 19 features which were selected out of the 30 features. In the next phase of the algorithm numerical labels were allotted to legitimate, suspicious and phishing websites i.e. 1, 0 and -1 respectively. Logistic regression (LR) and Support Vector Machine (SVM) was considered for implementation. Initially these two algorithms /classification techniques were implemented on complete dataset with 30 features and later then SVM and LR were applied on the reduced dataset having 19 features. Lastly this canonical step wise feature framework was compared using several performances measures like Precision /Recall/F-Measure and accuracy. Results were as follows.

Table no 06: Description of Dataset

Algorithm	Dataset	Precision	Recall	F-Measure	Accuracy
LR	Orig. DS (30)	91.50%	91.50%	91.50%	92.47%
LR	KMO Test (19)	89.07%	92.57%	90.78%	91.68%
SVM	Orig. DS (30)	94.95%	95.18%	95.06%	95.62%
SVM	KMO Test (19)	92.05%	93.63%	92.83%	93.59%

Detection of phishing attacks [29].In this study AntiPhishing simulator was developed to detect phishing sites along with this system it evaluates the keywords in URL and check the existing database and thus determines the content of the mail. The Preliminary stage is the data mining and the processing stage. The dataset used in this study is composed of the words that have the most used spam mail features today. Each word has its own weight. In the next stage Bayesian classifier was used to calculate the weight of the words and the spam word count was created. In this simulator text content is checked to determine whether the related message contains the phishing elements or not. By using the Bayesian classifier the scores are then added to the database. This structure also helps to add the feature of “add spam” to manage it on demand. This simulator can also create its own spam list to check whether the incoming mail is protected or not. Hence this simulator collects phishing and spam messages at a common point and allows to control the section of spam box where URL control feature examines the link in the address in the mail and classifies it as fake or genuine.

Email Phishing detection and prevention using data mining techniques [30].In this paper phishing detection method is proposed by using different machine learning and data mining techniques. This paper talks about different social engineering attacks and approaches targeted in these attacks. Different categories [31] are

1. Physical approach
2. Social approach and
3. Technical approach.

Some of the methods used were:-

1. Pretexting
2. Phishing
3. Spear phishing
4. Vishing
5. Baiting
6. Tailgating and
7. Email attachments etc

Technical phishing attacks include sending of suspicious emails / URLs and phishing websites. Phishing website detection takes help of URL structure to identify and classify websites and further for prevention mechanism various machine learning solutions like MLAPT technique can be used. Others include filtering the websites and checking the URL content with the Blacklist mechanism. In the proposed method, prevention mechanism is designed. It uses the concept of Information Gain to build a decision tree and WEKA tool is used for Data processing , classification , clustering and for data association features. In the dataset that is obtained for the study consist of 2949 suitable and harmless mails and 1370 phishing /data mails. The body and the header of the email are taken as the subject for study. Some of the features that were extracted includes the special keywords like $ sign, URL checking and other credentials like login passwords and links. At the next stage more than one data mining algorithms were applied which includes k-means, Naive Bayes and Neural network. It was found that the success rate of the decision tree classifier increased to ninety percent and Decision tree algorithm have also been implemented with J48 algorithm. The following results were obtained in different categories: - training set, supplied test set, cross validation and percentage split was found to be 82.45%, 89.98%, 53.85% and 78.54% TPR respectively.

III. Conclusion and Future work

On the phishing website dataset take from UCI repository when feature selection algorithm ARI i.e. Association rule was applied its features reduced from 31 to 27 and on this reduced dataset Lazy K Star classification algorithm produced the highest accuracy of 97.58%.When actual URL attributes were extracted from the chrome extension it showed that Random Forest classifier can prove to be useful in detecting the website as malicious or safe. Study also proved that 61% of the respondents were unaware of the phishing detection tools. C4.5 algorithm also proved to be beneficial when applied on phishing dataset which provided the accuracy of 82.6%.Experimental results also showed that PWCAC model obtained a good result in terms of F1 Score and accuracy. The performance of Random forest classifier was higher in terms of F1- Score, accuracy and AUC. In Future work, results can be tested with Big Data and also there is a scope for the development of optimized model which can achieve dimensionality reduction and can increase the performance parameters.

References

[1] R. Verma, N. Shashidhar, and N. Hossain, “Detecting phishing emails the natural language way,” in European Symposium on Research in Computer Security, pp. 824–841, Springer, 2012.

[2] Apwg, apwg report 2019. https://docs.apwg. org/reports/apwg_trends_report_q1_2019. pdf/.

[3] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, “Intelligent phishing website detection using random forest classifier,” in 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), pp. 1– 5, IEEE, 2017.

[4] M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” in 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–5, IEEE, 2018.

[5] M. Lichman et al., “Uci machine learning repository,” 2013.

[6] M. Thaker, M. Parikh, P. Shetty, V. Neogi, and S. Jaswal, “Detecting phishing websites using data mining,” in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1876–1879, IEEE, 2018.

[7] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, “Detection of phishing webpages based on visual similarity,” in Special interest tracks and posters of the 14th international conference on World Wide Web, pp. 1060–1061, ACM, 2005.

[8] G. Sonowal and K. Kuppusamy, “Phidma– a phishing detection model with multi-filter approach,” Journal of King Saud UniversityComputer and Information Sciences, 2017.

[9] P. K. Sahoo, “Data mining a way to solve phishing attacks,” in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1–5, IEEE, 2018.

[10] E. Kirda and C. Kruegel, “Protecting users against phishing attacks with antiphish,” in 29th Annual International Computer Software and Applications Conference (COMPSAC’05), vol. 1, pp. 517–524, IEEE, 2005.

[11] M. Wu, R. C. Miller, and G. Little, “Web wallet: preventing phishing attacks by revealing user intentions,” in Proceedings of the second symposium on Usable privacy and security, pp. 102–113, ACM, 2006.

[12] Y. Li, R. Xiao, J. Feng, and L. Zhao, “A semi-supervised learning approach for detection of phishing webpages,” Optik, vol. 124, no. 23, pp. 6027–6033, 2013.

[13] K.-P. Yee and K. Sitaker, “Passpet: convenient password management and phishing protection,” in Proceedings of the second symposium on Usable privacy and security, pp. 32–43, ACM, 2006.

[14] H. Sharma, E. Meenakshi, and S. K. Bhatia, “A comparative analysis and awareness survey of phishing detection tools,” in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1437–1442, IEEE, 2017.

[15] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, “Phinding phish: Evaluating anti-phishing tools,” 2006.

[16] T. Yue, J. Sun, and H. Chen, “Fine-grained mining and classification of malicious web pages,” in 2013 Fourth International Conference on Digital Manufacturing & Automation, pp. 616–619, IEEE, 2013.

[17] D. Canali, M. Cova, G. Vigna, and C. Kruegel, “Prophiler: a fast filter for the large-scale detection of malicious web pages,” in Proceedings of the 20th international conference on World wide web, pp. 197–206, ACM, 2011.

[18] A. Priya and E. Meenakshi, “Detection of phishing websites using c4. 5 data mining algorithm,” in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1468– 1472, IEEE, 2017.

[19] M. M. Al-Daeef, N. Basir, and M. M. Saudi, “A review of client-side toolbars as a user-oriented antiphishing solution,” in Advanced Computer and Communication Engineering Technology, pp. 427– 437, Springer, 2016.

[20] G. Aaron, “The state of phishing,” Computer Fraud & Security, vol. 2010, no. 6, pp. 5–8, 2010.

[21] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.
[22] G. Kostopoulos, Cyberspace and cybersecurity. Auerbach Publications, 2017.
[23] M. Alqahtani, “Phishing websites classification using association classification (pwcac),” in 2019 International Conference on Computer and Information Sciences (ICCIS), pp. 1–6, IEEE, 2019.
[24] M. J. Zaki and K. Gouda, “Fast vertical mining using diffsets,” in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 326–335, ACM, 2003.
[25] UCI, UCI Dataset. http://archive.ics.uci. edu/ml/datasets/Phishing+Websites.
[26] W. Fadheel, M. Abusharkh, and I. AbdelQader, “On feature selection for the prediction of phishing websites,” in 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), pp. 871–876, IEEE, 2017.
[27] A. A. Ahmed and N. A. Abdullah, “Real time detection of phishing websites,” in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1–6, IEEE, 2016.
[28] R. Mohammad, L. McCluskey, and F. Thabtah, “Uci machine learning repository: phishing websites data set (2015),” 2016.
[29] M. Baykara and Z. Z. G¨urel, “Detection of phishing attacks,” in 2018 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1– 5, IEEE, 2018.
[30] S¸. S¸ent¨urk, E. Yerli, and ˙I. So˘gukpınar, “Email phishing detection and prevention by using data mining techniques,” in 2017 International Conference on Computer Science and Engineering (UBMK), pp. 707–712, IEEE, 2017.
[31] K. Krombholz, H. Hobel, M. Huber, and E. Weippl, “Advanced social engineering attacks,” Journal of Information Security and applications, vol. 22, pp. 113–122, 2015.