IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Detection of Phishing URLs using Bayes Net and Naïve Bayes and evaluating the risk assessment using Attributable Risk

Priya Raj[1*], Meenakshi Mittal[2]

[1]Department of Computer Science & Technology (Cyber Security), Central University of Punjab, Bhatinda, India
[2]Department of Computer Science & Technology, Central University of Punjab, Bhatinda, India

**Corresponding: priyaraj9254@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 20/May/2018, Published: 31/May/2018

ABSTRACT- Phishing sites are manufactured or spurious URLs that are made by malignant people to imitate or imitate URLs of genuine URLs. An extensive bit of these sorts of URLs have most elevated twin to trap their casualties for tricks. Unwary Web customers may be successfully betrayed by this kind of trick. The effect is the break of data security through the exchange of private information and the losses may encounter the bad effects of financial losses and more example hacking. In this paper detection of phishing URLs is done by using Bayes net and Naïve Bayes algorithm and evaluation of risk regarding phishing URLs is done with the help of attributable risk. A training dataset of 1800 URLs (containing 1080 legitimate and 720 phished URLs) has been made to train the algorithms. Testing dataset of 720 URLs (containing 288 legitimate and 432 phished URLs) is used for making predictions using the DAG model classifier which is generated after the training of Bayes Net and Naïve Bayes Algorithm. True negative rate, True positive rate, false negative rate, false positive rate, Error rate and Accuracy are calculated after testing dataset by DAG classifier. Result shows Bayes Net has an accuracy of 71.3% and the Naïve Bayes has an accuracy of 80.5% and calculation of risk is done on the basis of attributable risk. If risk percentage for the URLs attributes is greater than 80% then risk is high, if it is between 50-80% then risk is medium and below 50% risk is low.

eywords- Attributable Risk, Bayes Net, Naïve Bayes, Phishing, Risk Assessment.

I. INTRODUCTION

Phishing sites are manufactured or spurious web pages that are made by malignant people to imitate or imitate web pages of genuine websites. An extensive bit of these sorts of Web pages have most elevated twin to trap their casualties for tricks. A portion of these sorts of Website pages look accurately or precisely like the certifiable ones. Unwary Web customers may be successfully betrayed by this kind of trick. Casualties of phishing Web pages may unveil their money related adjust, account points of interest, mystery scratch, Visa number, or other fundamental information to the phishing Site page proprietors. Phishing site are very complex issue to understand and to examine as many factors are involved. So identification of calculation is essential. Risk is the likelihood or probability of the event or acknowledgment of a risk.

Risk is the chance of the occurrence or recognition of a threat. There are three basic component of risk from an IT industries perspective i.e. asset, threat and vulnerability. So for resolving these issues there is a need of risk assessment. Here risk is calculated with the help of attributable risk. Attributable risk (AR) is calculated on the basis of phishing attributes present in the URLs, Attributable risk percent (AR %) is the percent of attributes present in the URLs. On the basis of the calculation, the risk is categorized into High, medium and Low.

A Bayes net or probabilistic directed acyclic graphical (DAG) model speaks to a gathering of random variable and their conditional dependencies with the assistance of DAG. Formally, Bayes net are DAGs which represent random variables as the nodes within the Bayesian network. Edges corresponds to the conditional dependencies from one node to other; variables that are conditionally independent of each other are represented by the nodes that are not connected. In BN every node of a graph is related to a probability function which can take, as an input, and set of values of nodes for parent variables, and generate the probability as output of the variable represented by the node.
Naïve Bayes classifier is direct technique which is used for constructing classifiers i.e. that classifies the model that assign class labels to the problem instances and

represented them as vector of feature values. All classifiers in naïve Bayes assume that the value of a specific feature is not dependent of the value of any other feature, given the class variable. The Naïve Bayes classifier works on the maximum likelihood. An advantage of naive Bayes is that it merely needs a small number of training data to estimate the parameters necessary for classification.

The paper is organized as follows, Section I contains the introduction of phishing website and algorithms i.e. Bayes Net and Naïve Bayes, Section II contain the related work regarding detection of phishing websites, Section III contain the methodology which is used to detect the phishing URLs, Section IV contain the result and discussion with calculation of risk for URLs, and Section V concludes the research work.

II. RELATED WORK

Phishing is one of the attracting strategies utilized by phishing specialists with the goal of abusing the individual points of interest of unsuspected users. Phishing should likewise be possible by means of messages which may contain connections to sites that appropriate malware. Phishing is consistently developing since it is anything but difficult to duplicate a whole site utilizing the HTML source code. Data mining procedures can enhance the assessment of phishing attacks. So identification of calculation is essential. Risk is the likelihood or probability of the event or acknowledgment of a risk. Identifying phishing website using a genetic algorithm and BP neural network which is a risk assessment model, proposed by Xiaoqian [6]. This approach is used to enhance the weights utilizing hereditary calculation and for edges of BP neural system to build up an information security risk assessment model, and for simulation MATLAB is used, predict the value of risk. The re-enactment comes about demonstrate the examination of GA-BP with the standard BP neural system, with better fitting impact and lower recreation mistake, is a good information security risk assessment model.

Another approach is proposed by Maher Aburrous et al which is a model based on the Fuzzy logic (FL) with combination of Data Mining algorithms for characterizing the e-banking phishing website factors and for investigating its techniques by classifying there phishing types and defining six e-banking phishing website attack criteria's with a layer structure. But finding the "right" feature set is a difficult problem and requires some intuition regarding the goal of data mining exercise [12].

Meenu Shukla and Sanjeev Sharma proposed strategy which is utilized to recognize phishing sites by utilizing URL features. It separates the essential features from URL and after that produced the outcome string with values representing to the URL behavior. Then perform WEKA test utilizing Random Forest Algorithm for various number of folds and compute the accuracy. It demonstrates a low false-positive rate and high accuracy of 97.31%. The proposed system can be utilized to give security and diminishing the harm caused by phishing attacks [14].About the social security network Annie Singla, Kamal Jain and Ajay Gairola influenced investigation of digital to space which is an inescapable and key place in each person's life now. With the approach of Internet trend, the quick digitization has occurred. In any case, with the fast digitization, there is a quick increment in the occurrence of cyber disasters. Disaster are estimated on two parameters – loss to property and life. In cyber disaster, there is a hell lot loss to money, zillions of individuals are getting influenced consistently and the loss of cash is in billions of dollars in each disaster. They are cutting down the economy of the country and individuals are likewise influencing on standard premise. Their protection is being endangered. Due to the Internet fad, E-trade, e-saving money, online gambling clubs are exploited. The weakness in the website page of Gmail is found out during research study. The studies of vulnerability lead to cyber disasters using Phishing attacks through which the passwords were recovered and gotten to of different individuals. Users should be more mindful and updated as far as innovation which can decrease these cyber disasters [15].

III. METHODOLOGY

The methodology is divided into following steps. Analysis of Bayes Net and Naïve Bayes algorithm is done by the Waikato Environment for Knowledge Analysis (WEKA) tool. Bayes Net and Naïve Bayes Algorithm generates a directed acyclic graph DAG after the evaluation of the training dataset.

The following steps which are used to analyze the Bayes Net and Naïve Bayes algorithm are as follows:

- Phishing URLs are collected from PhishTank.
- Features extraction of phishing websites.
- On the basis of extracted features create the training and testing dataset.
- Training of Bayes Net and Naïve Bayes algorithm using the training dataset and generate the directed acyclic graph classifier model.
- In the testing dataset prediction of the missing values in result attribute is done by using DAG classifier model for checking the URLs whether it is phished or not.
- Calculation of attributable risk on the basis of attributes, then categorized into high, medium, and low.

A. *Collection and extraction*

Collection of phished as well as legitimate URLs is done by using PhisTank. There are some features that differentiate the legitimate URLs from the phished or malicious URLs. There are thirty-one features for phishing URLs which are extracted i.e. are having IP address in the URL, usage of pop-up windows , having @ symbol in the URL, disabling of right click , redirection to new webpage, Long URL length, URL shortening, adding prefix and suffix separated by (-) to the domain, having sub domain and multi domain, HTTPS, Domain registration length, Favicon, Using Non-Standard Port, Existence of HTTPs token in the domain part of the URL, request URL, SFH, Link in <meta>,<script>,and<link> tags, URL of anchor, Submitting information to Email, Abnormal URL, Website forwarding, Status bar customization, Iframe redirection, Age of Domain, DNS record, Page Rank, Website Traffic, Google Index, Number of links pointing to page and Statics based report feature, etc.[13]

B. Creating datasets

After extraction of features of URLs creations of training and testing dataset is done. The training and testing dataset is saved as an arff file after its creation. The training dataset is used for the training of Bayes Net and Naïve Bayes algorithm and to create DAG classifier models. Pre-processing of training dataset need to be done first in WEKA. The training dataset of 1800 URLs is created in which 720 URLs are legitimate and 1080 URLs are phished. Test dataset of 720 in which 288 URLs are legitimate and 432 URLs are phished is created.

C. Training of Bayes Net and Naïve Bayes algorithm

The relation name "phishing" is used for the analysis of training dataset. 31 attributes are taken here. The attribute "Result" has two class values -1 for legitimate URLs and 1 for phished URLs. After loaded in WEKA the training dataset is preprocessed first. Bayes Net and Naïve Bayes algorithm are trained with the pre-processed training dataset, after that the DAG model is generated which is saved for prediction of URLs for the testing dataset.

D. Testing of Bayes Net and Naïve Bayes algorithm

A pre-processed trained DAG classifier model is used for making predictions in the testing process. DAG classifier is used for calculating the accuracy of testing dataset classifier where the missing values in the result attribute are presents in testing dataset. DAG classifier predicts the missing values on the basis of created confusion matrix. Earlier the DAG classifier model that was saved which needs to be loaded first after that, it is re-evaluated on the current test dataset. Here the testing dataset used contains 720 URLs. After testing the test dataset will get the accuracy of Bayes Net is 71.3% where correctly classified instances are 514 and accuracy of Naïve Bayes is 80.5% where correctly classified instances are 580.

IV. Results and discussion

After testing the data on the basis of pre-processed training dataset, will get the values of the parameters. Table 1 and Table 3 summaries the overall performance of Bayes Net and Naïve Bayes classifier in terms which are mentioned below. Table 2 and Table 4 shows the Confusion matrix of Bayes Net and Naïve Bayes classifier for the testing dataset of 720 URLs.

Correctly Classified Instances	514
Accuracy	0.713
Kappa statistic	0.4373
MAE	0.2867
RMSE	0.5343

Table 1. Performance of Bayes Net Classifier

	Predicted: No	Predicted: Yes
Actual: No	201	176
Actual: Yes	30	313

Table 2. Confusion matrix of Bayes Net

Analysis for Bayes Net: Bayes Net works on the conditional probability for each node in DAG. Here the root node which i.e. Result is conditionally depends on the children nodes which are the attributes. For e.g. Say the value of the result node is 1(phished) if the maximum occurrence of children nodes fulfilled the condition of a website for being phished. All 30 attributes have been selected for generating the graph as Bayes Net DAG needs all attributes to make predictions. In this DAG model there are 1 parent node which is result node and children nodes which are 30 attributes. The prediction of the result node is depend upon the conditional probability of each children node where children node contribute towards the prediction of a website whether it is legitimate or phished. On the basis of the generated DAG model after the evaluation of trained dataset will test the testing dataset and then prediction is made and confusion matrix is created. If the value of result node which is a parent node is -1 then the website is legitimate and if the value of result node is 1 then the website is phished.

Analysis of Naïve Bayes: Naïve Bayes classifier is based on the maximum likelihood or say maximum posteriori hypothesis and Bayes' Theorem. The naive assumes of class conditional independence to reduce the computational cost. It assumes that in a given class the effect of an attribute value is independent of the values of the other attributes which is known as class conditional independence. For e.g. say out of 30 attributes only 5 are

plays the major role of predicting the feature of website being phished.

Here taking this test URL **X** = (having_ip_address=1, URL_Length =1, having_At_Symbol= 1, Prefix_Suffix= 1, HTTPS_token= 1) where suppose these 5 attributes are present in the URL.

Then need to maximize P(X|Ci)P(Ci), where C is denoting the class.

P(Ci), the a priori probability of each class, can be estimated based on the training samples:

P(result = 1) = 1190/1800

P(result = -1) = 610/1800

To compute P(X|Ci) after which will know that the URL falls on which category, need to compute the following conditional probabilities:

Calculating the probability of having_ip_address when the Result=1, evaluate the training dataset 0f 1800 URLs will get 1045 rows where value of having_ip_address is 1 when simultaneously the value of result is 1which is 1190

P(having_ip_address = 1|Result = 1) = 1045/1190

Calculating the probability of having_ip_address when the Result = -1, evaluate the training dataset of 1800 URLs will get 363 rows where value of having_ip_address is 1 when simultaneously the value of result is -1which is 610

In same way, will calculate the same

P(having_ip_address = 1|Result = -1) = 363/610

P(URL_Length =1| Result = 1) = 336/1190

P(URL_Length=1|Result = -1) = 78/610

P(having_At_Symbol= 1|Result = 1) = 37/1190

P(having_At_Symbol= 1|Result = -1) = 1/610

P(Prefix_Suffix = 1|Result = 1) = 1190/1190

P(Prefix_Suffix = 1|Result = -1) = 608/610

P(HTTPS_token = 1|Result = 1) = 1141/1190

P(HTTPS_token = 1|Result = -1) = 310/610

Using the above probabilities, below will calculate the result when the value of result is 1

P(X|Result = 1) = P(having_ip_address = 1|Result = 1) P(URL_Length =1| Result = 1) P(having_At_Symbol= 1|Result = 1) P(Prefix_Suffix = 1|Result = 1) P(HTTPS_token = 1|Result = 1) =

$$= \frac{1045}{1190}\frac{336}{1190}\frac{37}{1190}\frac{1190}{1190}\frac{1141}{1190}$$

$$= 0.004$$

Similarly, will calculate the result when the value of result is -1

P(X|Result = -1)

$$= \frac{363}{610}\frac{78}{610}\frac{1}{610}\frac{608}{610}\frac{310}{610}$$

$$= 0.02$$

To find the class that maximizes P(X|Ci)P(Ci), Need to compute P(X|Result = 1)P(Result = 1) = 0.002 P(X|Result = -1)P(Result = -1) = 0.006

From here , will say that the URL X lie on class result = -1 which means that URL "X" is legitimate as Naïve Bayes works on maximum likelihood

Thus the naive Bayesian classifier predicts Result = -1 for sample X.

Correctly Classified Instances	580
Accuracy	0.805
Kappa statistic	0.5996
MAE	0.1956
RMSE	0.4404

Table 3. Performance of Naïve Bayes Classifier

	Predicted: No	Predicted: Yes
Actual: No	211	120
Actual: Yes	20	369

Table 4. Confusion matrix of Naïve Bayes

Overall Result:

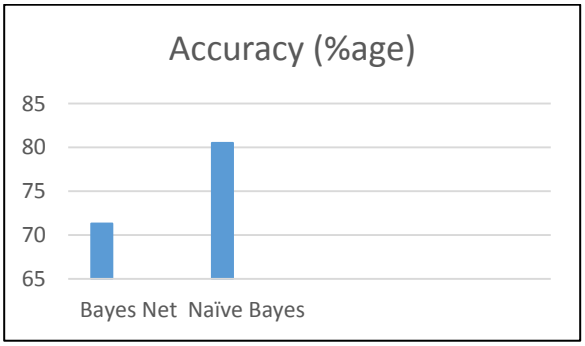

Fig. 1. Accuracy of Bayes Net and Naïve Bayes

Naïve Bayes gives the better result as compare to Bayes Net. Naïve Bayes works on the assumptions say all the features are conditionally independent of each other. Mostly this independence assumption works well for almost cases plus in Naïve Bayes works on the maximum likelihood whereas in Bayes Net there are not such assumptions, in Bayes Net all the features are conditionally dependent to each other.

Risk Assessment: After testing the URLs with the Bayes Net and Naïve Bayes, will calculate the risk of the URLs using attributable risk where attributable risk is calculated on the basis of the phishing attributes present in the URLs. The following table is shown below for some of the URLs.

Example: Say a URL having values, where '1'denotes phished, '0' denotes suspicious and '–1' denotes legitimate attribute.

1,-1,-1,-1,-1,1,1,-1,-1,1,0,1,-1,-1,-1,-1,-1,-1,1,1,0,0,-1,0,-1,-1,1,1,-1,-1,1,-1

As stated above, the attributable risk is calculated on the basis of number of 1's and 0's present in the URL. So here the 1 occur at 10 times and 0(zero) occurs 4 times, acc. to the formula

$$\frac{number\ of\ zeros+number\ of\ one's}{Total\ number\ of\ attributes} * 100 = \frac{10+4}{30} * 100 = 0.4666*100 = 46.6\%$$

Where 30 = total number of attributes

And 10 = number of 1's present in the URL

And 4 = number of 0's present in the URL

The percentage of risk is 46.6% in this URL that is less than 50% which means the URL has low risk and there is a very less probability of this URL being phished. The total number of URLs in the test dataset is 720 in which 254 URLs are having high risk, 12 URLs are having low risk and 454 URLs are having medium risk.

URLS	AR%	Category
http://recovery404.net/uk/log.php	60	Medium
http://outlookweb-upgrade.ukit.me	73	High
http://bit.ly/2gTY3Ef	50	Low
http://www.letshopmart.com/css/go.html	53	Medium
http://lava.hatchfactory.in/img/eng/ug/2c076ab56dc9991fe32557979946492c/?login=MANAGERZHOU@LONGZHOUTEX.COM&.verify?service=mail&data:text/html	80	High
http://beaversportsmedicine.com/i/Lloyds/	50	Low

V.CONCLUSION

In this research work weka is used for training and test datasets. Firstly analysis of features is done in the URLs for creating training dataset. Extraction of thirty features in URLs is done for generating the training and testing dataset. A training dataset of 1800 Urls and test datasets of 720 URLs is created. The created training dataset is used to train the Bayesian Networks i.e. Bayes Net, and Naïve Bayes algorithm in WEKA which generates a DAG classifier model. After this, accuracy of testing dataset is calculated. The result shows that the DAG model classifier of Bayes Net and Naïve Bayes can identify the URLs in the testing dataset with an accuracy of 71.3% and 81.3% respectively. Naive Bayes gives the better result as compare to Bayes Net as Naïve Bayes have an assumption of conditional independence between the features and uses maximum likelihood for computing the results. Risk assessment is calculated with the help of Attributable Risk which is calculated on the basis of the phishing attributes present in the URLs. After calculating the risk it is divided into 3 categories i.e. High (70% and above), medium (50% to 70%) and low (below 50%). The total number of URLs in the test dataset is 720 in which 254 URLs are having high risk, 12 URLs are having low risk and 454 URLs are having medium risk.

VI. References

[1]. B. K. Alese, O. Oyebade, O. A. Festus, O. Iyare, and A. F. Thompson, "Evaluation of information security risks using hybrid assessment model," The 9th International Conference for Internet Technology and Secured Transactions (ICITST-2014), pp. 387–395, 2014.

[2]. C.-T. Kuo, H.-M. Ruan, C.-L. Lei, and S.-J. Chen, "A Mechanism on Risk Analysis of Information Security with Dynamic Assessment," 2011 Third International Conference on Intelligent Networking and Collaborative Systems, pp. 643–646, 2011.

[3]. A. Tamjidyamcholo, "Information security risk reduction based on genetic algorithm," Proceedings Title: 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), pp. 122–127, 2012.

[4]. L. Zhou and Y. Zhou, "Gray relational analysis based method for information security risk assessment," 2012 7th International Conference on Computer Science & Education (ICCSE), pp. 1086–1089, 2012.

[5]. J. Bhattacharjee, A. Sengupta, and C. Mazumdar, "A formal methodology for Enterprise Information Security risk

assessment," 2013 International Conference on Risks and Security of Internet and Systems (CRiSIS), pp. 1–9, 2013.
[6]. X. Wu, Y. Shen, G. Zhang, and H. Zhi, "Information security risk assessment based on D-S evidence theory and improved TOPSIS," 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 153–156, 2016.
[7]. A. Fernandez and D. F. Garcia, "Complex vs. simple asset modeling approaches for information security risk assessment: Evaluation with MAGERIT methodology," 2016 Sixth International Conference on Innovative Computing Technology (INTECH), pp. 542–549, 2016.
[8]. G. Wangen, "Information Security Risk Assessment: A Method Comparison," Computer, vol. 50, no. 4, pp. 52–61, 2017.
[9]. S. Kondakci, "A causal model for information security risk assessment," 2010 Sixth International Conference on Information Assurance and Security, pp. 143–147, 2010.
[10]. J. Wang, K. Fan, W. Mo, and D. Xu, "A Method for Information Security Risk Assessment Based on the Dynamic Bayesian Network," 2016 International Conference on Networking and Network Applications (NaNA), 2016.
[11]. X. Chen, I. Bose, A. C. M. Leung, and C. Guo, "Assessing the severity of phishing attacks: A hybrid data mining approach," Decision Support Systems, vol. 50, no. 4, pp. 662–672, 2011.
[12]. M. R. Aburrous, A. Hossain, K. Dahal, and F. Thabatah, "Modelling Intelligent Phishing Detection System for E-banking Using Fuzzy Data Mining," 2009 International Conference on CyberWorlds, pp. 265–272, 2009.
[13]. R. M. Mohammad, L. Mccluskey, and F. Thabtah, "Intelligent rule-based phishing websites classification," IET Information Security, vol. 8, no. 3, pp. 153–160, Jan. 2014.
[14]. M. Shukla, S. Sharma "Analysis of Efficient Classification Algorithm for Detection of Phishing Site," International Journal of Scientific Research in Computer Science and Engineering, vol. 5, no. 3, pp. 136–141, Jun. 2017.
[15]. A. Singla, K. Jain, A. Gairola "Delving into Security of networks-Time's Ned," International Journal of Scientific Research in Network Security and Communication, pp. 1-8, Oct. 2014.

Authors Profile

Ms.Priya raj pursued B.Tech from Birla Institue of Applied Sciences in Computer Science and Technology in 2015. Currently she is pursuing M.tech in Computer Science &Technology (Cyber Security) fron Central University of Punjab.She is currently in his final year.Her main interested research area are Artifical Intelligence, Cryptography, and Automata.

Ms Meenakshi Mittal is pursed Master of Engineering in Computer Science from Punjab Engineering College University of Technology, Chandigarh, India.. She is currently working as Assistant Professor in Department of computer Science and Technology, Centeral University of Punjab, Bhatinda since 2011. She has published 6 research papers in conferences including IEEE and it's also available online. Her main research work focuses on Cryptography Algorithms, Mobile adhoc Networks, IOT, and Security. She has more than 6 years of teaching experience.

