International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

PHISH SAFE GURAD-Phishing Detection: Enhance Anti-Phishing System Using Machine Learning Algorithm

Vidya Mhaske-Dhamdhere, Sandeep Vanjale

Abstract: ***Today, a high usage of internet for online communication or online banking processing phishing is one of sensitive issues. Phisher can manage to gain users credentials and this causes financial loss of users.***

To detect phishing attack machine learning algorithms are used for classification. Anti-phishing system of PHISH SAFE-GUARD Phishing Detection (PSD-PD) based on URL features. To evaluate performance of proposed system 31 numbers of features are considered. System trained on 4601 phishing and legitimate URL with J48, Naïve Bayes, and Random forest. Our experiment shows result more than 92% accuracy in detecting phishing websites using J48 classifier.

Index Terms: Phishing websites and emails, Phishing Detection, classification, Machine Learning

I. INTRODUCTION

Today's world is online digital world. Every user user's internet for online communication and online banking processes for their daily life for his or her personal as well as professional work. While using online email or banking transaction services, users may share their credential or ignore security warning because of busy scheduled or unawareness of internet security measures. These reason causes phisher to trap to in phishing.

Phishing is one type of cybercrime or phishing attack, which leads to financial losses. There are number of techniques to design phishing attacks. The most popular thing is to create phishing website using contains of real one websites, which look like a same but their contents are replaced by small character in domain, subdomain, and no. of dots, '@', '-' like symbol replacement. It is very easy to replace these contents. Another way is send buck phishing emails or email embedded URLs are used. Phisher study user behavior and ask users to click on link. Email embedded link may redirected to fake website or it will ask user credentials. We need to detect such type of phishing attacks there are mainly two techniques: first one is user education and second is software detection approach.

User education approach is totally depending upon user behavior and knowledge on internet. We need to provide continues training. But in software detection approach blacklist is used for long back launch website detection. But phishing websites life is very short. It is very difficult to detect website using blacklist because we need to continue update blacklist then only zero day or Zero hour phishing attacks can be detected.so that we move to automated phishing attack detection using machine learning algorithms.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Mrs. Vidya Mhaske-Dhamdhere, PhD Research Scholar,Bharati Vidyapeeth Deeemed to be University, Pune. India
Dr. Sandeep Vanjale, Professor in Computer Engg. Dept. Bharati Vidyapeeth Deeemed to be University, Pune. India

II . LITERATURE SURVEY

Mehdi Bagnoli [1] has implemented nonlinear regression method for phishing website detection. From UCI data set 20 features are extracted using method of decision tree and wrapper accuracy is 96.32%. These feature selection method remove unwanted or less useful features because of that reduces training time of data. Nonlinear regression is used to find the functional relationship between input and output, which is mainly used for feature selection. This proposed harmony search method is implemented in Weka 3.6 and compared with SVM. Harmony search gives 92.80% and SVM gives 91.83% accuracy.

Hossein Abroshan, Jan Devos [2] discusses root causes of phishing attacks. There are number of Anti-phishing tools and techniques availed. But users are not aware of anti-phishing techniques or they are aware but not used for security purposed. Because of this reason phishers use user behavior to design phishing attacks. So that every time new types of tricks and tactics are used by phisher.it will very difficult for user to identify attack types also it will take more time. For overcome this cause's user behavior monitor on online activity and anti-phishing training is needed. Anti-phishing techniques includes phishing website, email, network detection approached are required. Heuristics and machine learning methods are used for web based feature selection for phishing websites detection but their complexity is high. In this paper author has suggested to develop and to prevent from reducing phishing attacks.

Mohammad Karim Sohrabi, Firoozeh Karimi [3] has online spam detection design and implemented spam detection on Facebook using feature selection. Generally spammer redirects users to malicious pages posted by users or URL. Spammer design messages on basis of user's behavior and social relationship information. In this wrapper based and filter based approaches are used. In filter based approach, features of dataset evaluated and select subset of features without using any learning approach. In wrapper based approach uses classification techniques to select best features. Online spam detection feature selection is plays important role; it will reduce training time and increase performance.

Particle swarm optimization is used in proposed system for feature selection and clustering SVM and decision tree is used. SVM has best performance but decision tree has 87% detection accuracy but time complexity is less. If we consider time complexity for system evaluation criteria decision tree is best with threshold value 0.6.

Zhag Yuqiang, He Jingsha, XU Jing [4] has design new anti-spam model for phishing .to deal with spam email mainly junk email problem caused by the email address leakage for a majority of internet users. Author has design new privacy protection model to solve junk email problem. This spam filter problem give 100% spam filtered scheme.

Shafi Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Psho [5] has done spam email detection comparative analysis. Junk emails are sending to multiple users at a time. Machine learning algorithms of Bayesian Logistic Regression, Hidden Naïve Bayes, and Radial Basis Function are used on Weka tool. Spam base dataset is used with 57 features on 4601 data instances for experiment with 10 fold cross validation and 66% split. Performance is compared on basis of accuracy, precision, recall, f-measure, Root mean squared error. Rotation forest gives 94.2% accuracy, which is best.

Aniket Kumar Jain, B. B. Gupta [6] had implemented proposed system using machine learning algorithm logical regression using feature hyperlinks in email. For phishing classification 2544 dataset size with 12 features are consider, which gives 98.39% true positive rate and accuracy 98.42%.phishing detection totally depend on source code of websites.in this paper authors has used source code for website phishing detection ,it is time consuming process.

Anikit Kumar Jain, B.B. Gupta [7] has design anti-phishing system of PHISH-SAFE using machine learning algorithm.in this anti-phishing system URL features based phishing detection done. Total 14 features from URL as consider detecting phishing website. Total 33,000 URL taken form phish tank dataset and then trained with SVM and Naïve Bayes classifiers. SVM gives 90% accuracy which is more than Naïve Bayes.

Hassan Y.A. Abutair, Abdelfettah Belghith [8] has implemented case based reasoning (CBR) system apply for phishing detection. Only 572 URLs are used. CBR works in 4 phase retrieve, reuse, revise, retain and wrapper based feature selection is used for offline and online experiments with accuracy from 95.62% to 98.0%.this method predict zero hour phishing attacks easily.

Vidya Mhaske-Dhamdhere, Dr. Sandeep Vanjale [9] has design user awareness training for user education. Initially test that users having knowledge of phishing websites or not. Some questionaries' are design. And send to approximately 8000 users from different field .author got approximately 5000 feedback. Most of users fall into phishing because of lack of knowledge .so authors has decided to give user phishing email awareness training. For training user those are having less internet usages experience as well those are having good computer knowledge with age group 18-25 years old users are consider for user education. Mainly author has taken computer engineering student whose age group is from 18-25 years old. This training is divided into 3 parts .first one is before training some phishing and real email send to user and monitor their performance. Very less users are able to identify phishing email or they click on email embedded link and share their personal credential. Then author give training those student and explain does and don't this was done in second part.in third part author again design phishing and real email and send to trained users and check performance. Author has compared first and third part performance only 10% improvement. So author reach conclusion that continues training of user awareness is required. Vidya Mhaske-Dhamdhere, Dr. Sandeep Vanjale [10] author has used dataset and clssify dataset using machine learning algorithm, which gives good accuracy. Anandita, Dhiendra Yadav, Priyanka Paliwal, Divya Kumar, Rejesh Tripathi 11] has ensemble 5 machine learning algorithm. Ensemble system take input from 5 machine learning algorithm Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Random forest classifier, K-Nearest Neighbors, Support Vector Machines has work in three phases-preprocessing of dataset, feature extraction, classification. Total 15 features are taken form preprocessing of email header and body part compare with 2550 size of Spam Asian dataset. Random Forest give 94.0% accuracy which is highest than other algorithm which is used in ensemble system.

III PROPOSED SYSTEM

Anti-phishing system of PHISH SAFE-GUARD Phishing Detection (PSD-PD) based on URL features. PSD-PD has three phases: feature selection, feature extraction, classification. Input to the machine learning algorithm is URL. Foe this Spam base dataset is used with 5000 instant ace with 31 features of URL. Proposed system implemented in java and imported WEKA tool for dataset analysis.First phase is URL feature selection done using wrapper based approach for best feature selection. Then features and extracted from URL, then apply machine learning algorithm for classification.

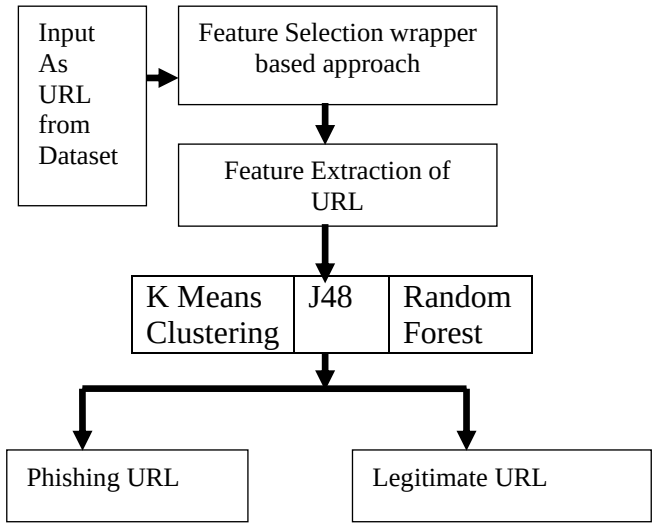

Figure1.PHISH-SAFE-GUARDPhishingDetection (PSD-PD).

IV RESULT

CSDMC2010 dataset is used for with 4327 dataset instances apply on machine learning algorithm for classification phishing emails. We apply dataset on Naïve Bayes, J48, and Random Forest. J48 gives 92% true positive which is higher than other algorithm. J48 has true positive 92.0% but random forest has highest precision that is 95.50%, which is shown in table 1[10] and figure2.

Table1.Machine learning algorithm result comparison

Dataset name	Spam base	Corpus CSDMC2010	Spam
Dataset size	4601	4327	2701
Feature /attributes	58	38	31
Phishing	2788(61%)	2949(77%)	1193(45%)
legitimate	1813(39%)	1378(23%)	1477(55%)

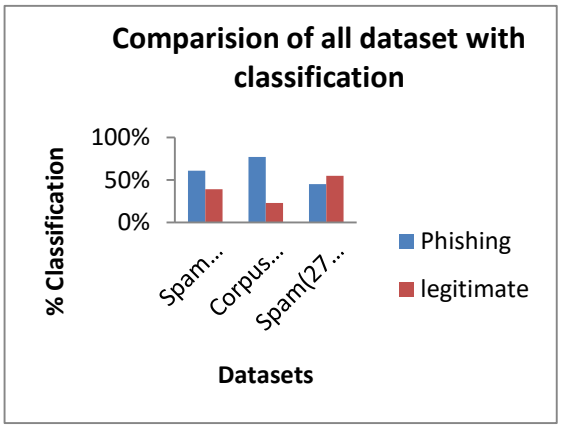

Figure 2. Classification of results with comparison of parameters

Table 2. Comparison of datasets with classification

parameter	Random Forest	J48	Naïve Bayes
	Phishing URL	Phishing URL	Phishing URL
True positive	91.20%	92.00%	49.00%
False Negative	0.025	0.56%	3.10%
Precision	95.50%	91.30%	66.60%
Recall	91.20%	94.20%	95.10%

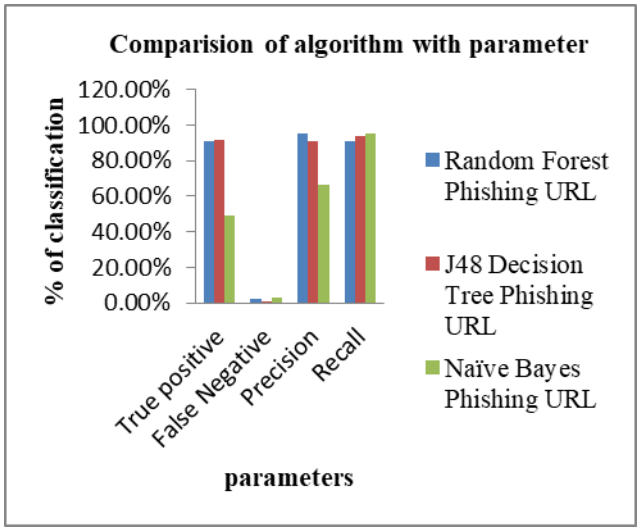

Figure 3Comparision of datasets with classification

Second experiment performed on different dataset like spam base CSDM2010 and spam with data instances 4601, 4327, 2701 on Naïve Bayes algorithm. If we take maximum features for checking performance it will take more times because some feature may be not irrevenlent.

V . CONCLUSION

Phishing detection is one of challenging task. Software detection approach plays important role in phishing detection apply on machine learning algorithm. But each algorithm has its own advantage and disadvantage.so we combine 3 algorithms to check it performance in PHISH SAFE-GUARD Phishing Detection (PSD-PD) on CSDMC2010 dataset with 4327 instances with 38 features on J48, Random Forest, Naïve Bayes. But Random forest perform better with accuracy 92%.

In future we need to increase data size but number of future keeping constant and check the performance of algorithm in term of classification and time required for execution.

REFERENCES

1. Mehdi Bagnoli, Mohammad pour Mahmood, Aghababa Vahid solauk,"Heuristic nonlinear regression strategy for detecting phishing websites", Soft Computing, springer 2018,
2. Hossein Abroshan, Jan Devos,"Phishing attacks root causes", crisis 2017, Srpinger2018.
3. Mohammad Karim Sohrabi, Firoozeh Karimi,"A feature selection approach to detect spam in the Facebook social network", Arab journals of science and engineering, springer 2017.
4. Zhag Yuqiang, He Jingsha, XU Jing," A New anti- phishing model based on email address concealment techniques", Wuhan University Journal of Natural Science,Vol.28,No.1,2018 PP.79-83
5. Shafi Muhammad Abdulhamid, Maryam Shuaib, Oluwafemi Psho," Comparative analysis of classification algorithms for email spam detection", I.J. Computer Network and Information Security, MECS-2018, PP. 60-67.
6. Aniket Kumar Jain, B. B. Gupta," A machine learning based approach for phishing detection using hyperlinks information", Journal of Ambient Intelligent and Humanized ocxmputing,Springer2018.
7. Anikit kumar Jain, B.B. Gupta ," PHISH –SAFE: URL feature based phishing detection system using machine learning, Cyber Security, Advances in Intelligent Systems and Computing, springer 2018,pp.467-474.

8. Hassan Y.A. Abutair, Abdelfettah Belghith," Using Case Based reasoning for phishing Detection ",8 th international conference eon Ambient System, networks and technologies, Elsevier-2017 , pp. 281-288.
9. Vidya Mhaske-Dhamdhere , Dr. Sandeep Vanjale," A novel approach for phishing emails real time classification using k-means algorithm", International Journal of Engineering & Technology, 7 (1.2) (2018) 96-100
10. Vidya Mhaske-Dhamdhere , Dr.Sandeep Vanjal, "Phishing emails classification and clustering using data mining algorithm" Vol 8, No 6,pp, 5326-5332
11. Anandita, Dhiendra Yadav, Priyanka Paliwal, Divya Kumar, Rajesh Tripathi " A Novel Ensemble Based identification of Phishing Emails.",ICMLC2- ACM conference 2017

AUTHORS PROFILE

Ms. Vidya Mhaske-Dhamdhere is Ph.D. student in the Bharati Vidyapeeth Deemed to be University, Pune. She has over 10 years of academic experience. Her interest area of research interest is Cyber security using Machine learning and data mining

Dr. Sandeep Vanjale received Ph.D. in Computer Science from BVDUCOEP. He has over 20 years of academic. His research interests are in Network security, WSN, Cyber Security.