International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.5 November – 2021 DOI: 10.46647/ijetms.2021.v05i06.003
ISSN: 2581-4621

Prediction of Phishing Website for e-Banking Using Data Mining Techniques

Nemmi Swathi[1], C Maddilety[2]

[1]*M.Tech Student, Dept. of CSE, Dr.K.V.Subba Reddy College of Engineering for Women, Kurnool, A.P*

[2]*Assistant Professor, Dept. of CSE, Dr.K.V.Subba Reddy College of Engineering for Women, Kurnool, A.P*

[1]swathiyadav.123@gmail.com
[2]chinna.mdh@gmail.com

***Abstract*— Classification Data Mining (DM) Techniques can be a very useful tool in detecting and identifying e-banking phishing websites. In this paper, we present a novel approach to overcome the difficulty and complexity in detecting and predicting e-banking phishing website. We proposed an intelligent resilient and effective model that is based on using association and classification Data Mining algorithms. These algorithms were used to characterize and identify all the factors and rules in order to classify the phishing website and the relationship that correlate them with each other. We implemented six different classification algorithm and techniques to extract the phishing training data sets criteria to classify their legitimacy. We also compared their performances, accuracy, number of rules generated and speed. A Phishing Case study was applied to illustrate the website phishing process. The rules generated from the associative classification model showed the relationship between some important characteristics like URL and Domain Identity, and Security and Encryption criteria in the final phishing detection rate. The experimental results demonstrated the feasibility of using Associative Classification techniques in real applications and its better performance as compared to other traditional classifications algorithms.**

***Keywords*— Phishing, Fuzzy Logic, Data Mining, Classification, Association, e-banking risk assessment.**

I. INTRODUCTION

There are number of users who purchase products online and make payment through e- banking. There are e- banking websites who ask user to provide sensitive data such as username, password or credit card details etc often for malicious reasons. This type of e-banking websites is known as phishing website. In order to detect and predict e-banking phishing website. We proposed an intelligent, flexible and effective system that is based on using classification Data mining algorithm. We implemented classification algorithm and techniques to extract the phishing data sets criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate. Once user makes transaction through online when he makes payment through e-banking website our system will use data mining algorithm to detect whether the e-banking website is phishing website or not. This application can be used by many E-commerce enterprises in order to make the whole transaction process secure. Data mining algorithm used in this system provides better performance as compared to other traditional classifications algorithms. With the help of this system user can also purchase products online without any hesitation.

Phishing websites is a semantic attack which targets the user rather than the computer. It is a relatively new Internet crime in comparison with other forms, e.g., virus and hacking. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. The word phishing from the phrase "website phishing" is a variation on the word "fishing". The idea is that bait is thrown out with the hopes that a user will grab it and bite into it just like the fish. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites [7]. The motivation behind this study is to create a resilient and effective method that uses Data Mining algorithms and tools to detect e-banking phishing websites in an Artificial Intelligent technique. Associative and classification algorithms can be very useful in predicting Phishing websites. It can give us answers about what are the most important e-banking phishing website characteristics and indicators and how they relate with each other. Comparing between different Data Mining classification and association methods and techniques is also a goal of this investigation since there are only few studies that compares different data mining techniques in predicting phishing websites.

II.SYSTEM ANALYSIS

Phishing Characteristics and Indicators

we managed to gather 27 phishing features and indicators and clustered them into six Criteria (URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor), and each criteria has its own phishing components. For example, URL & Domain Identity Criteria has five phishing indicator components (Using IP address, abnormal request URL, abnormal URL of anchor, abnormal DNS record and abnormal URL).

The full list is shown in table 3 which is used later on our analysis and methodology study.

Data Mining has been described as "the nontrivial extraction of implicit, unknown, and potentially useful information from large

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.5 November – 2021 DOI: 10.46647/ijetms.2021.v05i06.003
ISSN: 2581-4621

data sets [18]. It is a powerful new technology to help researchers focus on the important information in their data archive. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [19].

Despite growing efforts to educate users and create better detection tools, users are still very susceptible to phishing attacks. Unfortunately, due to the nature of the attacks, it is very difficult to estimate the number of people who actually fall victim.

Table1.Main Phishing Indicators with Criteria

Criteria	N	Phishing Indicators
URL & Domain Identity	1	Using IP address
	2	Abnormal request URL
	3	Abnormal URL of anchor
	4	Abnormal DNS record
	5	Abnormal URL
Security & Encryption	1	Using SSL certificate (Padlock Icon)
	2	Certificate authority
	3	Abnormal cookie
	4	Distinguished names certificate
Source Code & Java script	1	Redirect pages
	2	Straddling attack
	3	Pharming attack
	4	OnMouseOver to hide the Link
Page Style & Contents	1	Spelling errors
	2	Copying website
	3	Using forms with *Submit* button
	4	Using pop-ups windows
	5	Disabling right-click
Web Address Bar	1	Long URL address
	2	Replacing similar char for URL
	3	Adding a prefix or suffix
	4	Using the @ Symbol to confuse
Social Human Factor	1	Emphasis on security
	2	Public generic salutation
	3	Buying time to access accounts

III.SYSTEM CONSTRUCTION

CBA and MCAR experiments were conducted using an implementation version provided by the authors of [20], [26]. We have chosen these algorithms based on the different strategies they use to generate the rules and since their learnt classifiers are easily understood by human.

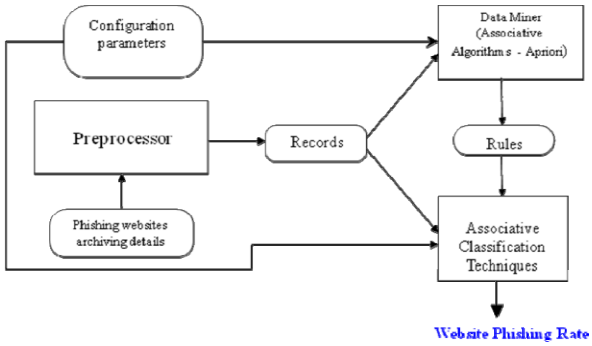

Figure 1. AC Model for Detecting Phishing Websites

We used two web access archives, one from APWG archive [1] and one from Phishtank archive [24]. We managed to extract the whole 27 phishing security features and indicators and clustered them to its 6 corresponding criteria as mentioned before in table 1.

Mining e-Banking Phishing Challenges

The age of the dataset is the most significant problem, which is particularly relevant with the phishing corpus. E- banking Phishing websites are short-lived, often lasting only in the order of 48 hours. Some of our features can therefore not be extracted from older websites, making our tests difficult. The average phishing site stays live for approximately 2.25 days [14]. Furthermore, the process of transforming the original e-banking phishing website archives into record feature row data sets is not without error. It requires the use of heuristics at several steps. Thus high accuracy from the data mining algorithms cannot be expected. However, the evidence supporting the golden nuggets comes from a number different algorithms and feature sets and we believe it is compelling [15].

Classification Algorithm

We utilize six different common DM classification algorithms (C4.5, JRip, PART, PRISM, CBA and MCAR). Our choice of these methods is based on the different strategies they used in learning rules from data sets [17]. The C4.5 algorithm [23] employs divide and conquer approach, and the RIPPER algorithm uses separate and conquer approach. The choice of PART algorithm is based on the fact that it combines both approaches to generate a set of rules. It adapts separate- and-conquer to generate a set of rules and uses divide- and-conquer to build partial decision trees. PRISM is a classification rule which can only deal with nominal attributes and doesn't do any pruning. It implements a top- down (general to specific) sequential-covering algorithm that employs a simple accuracy-based metric to pick an appropriate rule antecedent during rule construction. CBA algorithm employs association rule mining [20] to learn the classifier and then adds a pruning and prediction steps. Finally, MCAR algorithm consists of two phases: rules generation and a classifier builder. In the first phase, MCAR scans the training data set to discover frequent single items, and then recursively combines the items generated to produce items involving more attributes. MCAR then generates ranks and stores the rules. In the

Website: ijetms.in Issue: 6 Volume No.5 November – 2021 DOI: 10.46647/ijetms.2021.v05i06.003
ISSN: 2581-4621

second phase, the rules are used to generate a classifier by considering their effectiveness on the training data set. This results in a classification approach named associative classification [6] [7]. MCAR utilizes database coverage pruning to decrease the number of rules. Since without adding constraints on the rule discovery, the very large numbers of rules, make humans unable to understand classifier. This pruning technique tests the generated rules against the training data set, and only high quality rules that cover at least one training instance not considered by other higher ranked rules are kept for later classification.

MCAR (*Multi Class Classification based on Association Rule*) algorithm which had an accuracy of 88.4 % and error rate of 12.622 % is shown below.

Rule 1:

Social_Human_Factor = Fraud

Web_Address_Bar = Fraud

Page_Style_&_Contents = Doubtful

->class = Phishing

Rule 16:

Web_Address_Bar = Genuine

Security_&_Encryption = Doubtful

URL_Domain_Identity = Doubtful

->class = Legitimate

Rule 22:

Social_Human_Factor = Genuine

Page_Style_&_Contents = Doubtful

->class = Suspicious

We recorded the prediction accuracy and the number of rules generated by the traditional classification algorithms and the new associative classification approaches we used in Table 5 and 6 respectively. Experiments were conducted using stratified ten-fold cross-validation.

Table 5. Results From Weka four Classifiers

	C4.5	P.A.R.T.	JRip	PRISM
Test Mode	10 FOLD CROSS VALIDATION			
Attributes	Url Domain Identity Security & Encryption Source Code & Java Page Style & Contents Web Address Bar Social Human Factor			
No.of Rules	57	38	14	155
Correctly	848	869	818	855
Classified	(84.2 %)	(86.3 %)	(81.3%)	(84.9%)
Incorrectly	158	137	188	141
Classified	(15.7%)	(13.6%)	(18.6 %)	(14.0%)
Instances	1006	1006	1006	1006

Table 6. Results from CBA and MCAR Classifiers

	CBA	MCAR
Num of Test Case	1006	1006
Correct Prediction	873	886
Error Rate	13.452%	12.622%
Min Sup	20.000%	20.000%
Min Conf	100.000%	100.000%
Rule Limit	80000	80000
Level Limit	6	6
Number of rules	15	22

IV.CONCLUSION

The most important way to protect the user from phishing attack is the education awareness. Internet users must be aware of all security tips, which are given by experts. Every user should also be trained not to blindly follow the links to websites where they have to enter their sensitive information. It is essential to check the URL before entering the website. It has become a serious network security problem, facing financial loss of billions of dollars to both consumers and the e-commerce companies. And perhaps more eventually, phishing has made e-commerce distrusted and attractive to normal consumers.

E-banking phishing website model based on classification data mining showed the significance importance of the phishing website two criteria's (URL & Domain Identity) and (Security & Encryption) in the final phishing detection rate, and also showed the insignificant trivial influence of some other criteria like 'Page Style & content' and 'Social Human Factor' in the final phishing rate. The rules generated from the associative classification model showed the correlation and relationship between some of their characteristics which can help us in building phishing website detection system. The experiments demonstrate the feasibility of using Associative Classification techniques in real applications involving large databases and its better performance as compared to other traditional classification algorithms.

In Future System can upgrade to automatic Detect the web page and the compatibility of the Application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. Phish Checker application also can be upgraded into the web phone application in detecting phishing on the mobile platform.

V.REFERENCES

[1] Jian Mao, Wenqian Tian, Pei Li, Tao Wei, Zhenkai Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity", IEEE Access Volume: 5

[2] Zou Futai, Gang Yuxiang, Pei Bei, Pan Li, Li Linsen, "Web Phishing Detection Based on Graph Mining", 2016 2nd IEEE International Conference on Computer and Communications (ICCC)

Website: ijetms.in Issue: 6 Volume No.5 November – 2021 DOI: 10.46647/ijetms.2021.v05i06.003
ISSN: 2581-4621

[3] Nick Williams, Shujun Li, "Simulating human detection of phishing websites: An investigation into the applicability of ACT-R cognitive behaviour architecture model",

[4]Xin Mei Choo, Kang Leng Chiew,Dayang Hanani Abang Ibrahim,Nadianatra Musa, San Nah Sze, Wei King Tiong, "Feature-Based Phishing Detection Technique", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)

[5] Giovanni Armano, Samuel Marchal and N. Asokan, "Real-Time Client-Side Phishing Prevention Add-on", 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)

[6] Trupti A. Kumbhare and Prof. Santosh V. Chobe, "An Overview of Association Rule Mining Algorithms",

[7] S.Neelamegam, Dr.E.Ramaraj, "Classification algorithm in Data mining: An Overview", International Journal of P2P Network Trends and Technology (IJPTT) - Volume 3 Issue 5 September to October 2013

[8] Varsharani Ramdas Hawanna, V. Y. Kulkarni and R. A. Rane "A Novel Algorithm to Detect Phishing URLs.", 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)

[9] Jun Hu, Xiangzhu Zhang,Yuchun Ji, Hanbing Yan, Li Ding, Jia Li and Huiming Meng "Detecting Phishing Websites Based on the Study of the Financial Industry Webserver Logs." , 2016 3rd International Conference on Information Science and Control Engineering (ICISCE)

[10] Jian Mao, Wenqian Tian, Pei Li, Tao Wei, Zhenkai Liang, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity", IEEE Access Volume: 5

[11] Web Phishing Detection Using a Deep Learning Framework. Hindawi Wireless Communications and Mobile Computing Volume 2018, Article ID 4678746, 9 pages.

[12] Phishing Websites Detection Using Machine Learning R. Kiruthiga, D. Akila. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019

[13] Detection of URL based Phishing Attacks using Machine Learning. Published by: International Journal of Engineering Research & Technology (IJERT) http://www.ijert.orgISSN:2278-0181 Vol. 8 Issue 11, November-2019R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev. in press.

[14] FDIC, Tech. Rep., "Putting an end to account-hijacking identity theft", Dec.2004.[Online].Available:http://www.fdic.gov/consumers/idtheftstudy/identity theft.pdf.

[15] Ian Fette, Norman Sadeh and Anthony Tomasic, "Learning to Detect Phishing Emails", Institute for Software Research

International, CMU-ISRI-06-112, June 2006.

[16] WEKA - University of Waikato, New Zealand, EN, 2006: "Weka - Data Mining with Open Source Machine Learning Software inJava";http://www.cs.waikato.ac.nz/ ml/weka (2006/01/31).

[17] Sebastian Misch, "Content Negotiatio in Internet Mail", Diploma Thesis, University of Applied Sciences Cologne, Mat.No.: 7042524, February 2006.

[18] Kantardzic and Mehmed. "*Data Mining: Concepts, Models, Methods, and Algorithms*.", John Wiley & Sons. ISBN 0471228524. OCLC 50055336, 2003.

[19] U.M. Fayyad, "Mining Databases: Towards Algorithms for Discovery," Data Eng. Bull., vol. 21, no. 1, pp. 39-48, 1998.

[20] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining." *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98,*

Plenary Presentation), New York, USA, 1998.

[21] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann,

San Francisco, CA, 2005