IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-5 E-ISSN: 2347-2693

A Comparative Study of Existing Data Mining Techniques for Phishing Detection

M. Shukla[1*], S. Sharma[2]

[1]Dept. of CSE and IT, Madhav Institute of Technology and Science (RGPV University), Gwalior, India
[2]Dept. of CSE and IT, Madhav Institute of Technology and Science (RGPV University), Gwalior, India

[*]*Corresponding Author: shuklameenu03@gmail.com, Mob.: +91-9109501423*

Available online at: www.ijcseonline.org

Received: 11/Apr/2017, Revised: 23/Apr/2017, Accepted: 20/May/2017, Published: 30/May/2017

***Abstract*—** Nowadays phishing become a major threat on internet. Phishing is a kind of attack for defacement of website in which attacker can access sensitive information of users. Phishers are one who create website same as the trusted website with the same content and designs of the trusted website. Phishing can be done through email, websites and malicious software to get intellectual information, business secrets or military secrets etc. This paper is explored the various researches for avoiding phishing and detecting phishing symptoms. Many researchers have been proposed various methods for algorithms for avoiding all conditions with the detection of phishing using data mining techniques so that any user can use internet effectively. This paper is based on Associative Classification methods of data mining for avoidance of phishing attack.

***Keywords*—** Phishing, Associative Classification, Data Mining, Avoidance methods of phishing

I. INTRODUCTION

A main security threat to online business comes from what becomes to be referred to as "Phishing Attacks". In such attacks, malicious people create webpages that mimic the webpages of legitimate websites. Clients of the legitimate website falsely access the faked site and render their financial and personal information to malicious people whom might use this information to perform unlawful and criminal activities. Such criminal act causes a lot of loss for both the clients and the legitimate companies [1]. Phisher typically create web pages that are visually very similar to the real web pages in order to cheat their victims. An unaware client might be easily deceived by this kind of scam. The Targets of a phishing Web page may disclose their bank account, password, credit card number, or any other personal information to the phishing Web page owners. While phishing is a relatively new Internet crime when compared to other forms (e.g., viruses and hacking), a recognizable increase in the number and severity of phishing attacks is reported (Anti-Phishing Working Group, 2011).

According to the Anti-Phishing Working Group (APWG) reports for the Q4 on 2012 (Activity, 2012), the APGW received reports of 28,195 unique phishing sites in December. During Q4, about 30% of personal computers worldwide were infected with malware. Indeed, financial services found to be the most-targeted industry sector in the Q4 of 2012. Moreover, Payment Services eclipsed retail/services have the second-highest industry sector for targeted attacks.

The purpose of the phishing website is to steal the targets' personal information by visiting and surfing a false webpage that looks like a true one of a legitimate bank or company and asks the target to enter their personal information such as their username, account number, password, credit card number, etc. The impact is the break of information security through the compromise of confidential data and the victims may finally suffer losses of money or other kinds of assets. The attackers might also commit identity theft crimes using the victim's stolen information. Moreover, phishing attacks also damage the reputation of the attacked financial institutes since customers become less confident that they can securely access their accounts. Therefore, they might switch to other institutes. Phishing [2] has a massive negative effect on organizations' incomes, customer relationships, business marketing, and overall corporate field. Phishing attacks [3] may cost companies hundreds of thousands of dollars per attack in fraud-related losses and personnel time. One of the new data mining techniques is associative classification (AC)

which integrates two known data mining tasks, association rule mining as well as classification. The classification step is added in order to use the produced classifier model for the purpose of prediction. The two data mining tasks [1] are analogues, with the exception that classification aims to forecast the class label, while association rule describes correlations among items in a transactional dataset, several studies provided evidences that AC usually extracts better classifiers with reference to classification accuracy than other traditional classification approaches, such as decision trees, and rule induction.

Data mining provides a new solution to detect phishing issue. That is why data mining is an immerging research trend towards the detecting and preventing phishing website. Associative Classification is a grooming research method in data mining. Therefore it is an interesting research topic for detecting phishing using associative classification. Associative Classification (AC) [5] in data mining is one of the promising approaches that can make use of the features extracted from phishing and legitimate websites to find patterns among them. This approach normally devises classifiers (set of rules) that are simple yet accurate. The decision-making process becomes reliable because these decisions are made based on rules discovered from historical data by the AC algorithm. Although plenty of applications are available for combating phishing websites few of them make use of AC data. Phishing is a typical classification problem in which the goal is to assign a test data (a new website) one of the predefined classes (phishy, legitimate, suspicious, etc.). Once a website is loaded on the browser a set of feature values will be extracted. Those features have a strong influence in determining the website type by applying the rules that have been previously found by the AC algorithm from the historical data (already labeled websites). Then, the chosen rule's class will be assigned to the browsed website and an appropriate action will take place. For instance, a message or an alarm will be fired to alert the user of the risk. In this paper, the problem of phishing detection is explored using AC approach in data mining. Gupta et al. [5] primarily test a developed AC algorithm called MCAC and compare it with other AC and rule induction algorithms on phishing data.

II. ASSOCIATIVE CLASSIFICATION MODEL

Basically, the data mining classification and association rule approaches are utilized in a e-banking phishing website detection model, to find the most important phishing features and significant patterns of phishing characteristic or factors in the e-banking phishing website archive data. In fig 1, the Configuration Parameters uses two basic approaches i.e., Data Mining using associative classification algorithm and Associative Classification techniques, in which the processed achieved phishy websites details are fetched in the form of records and then the decision is made accordingly with the help of predefined techniques.

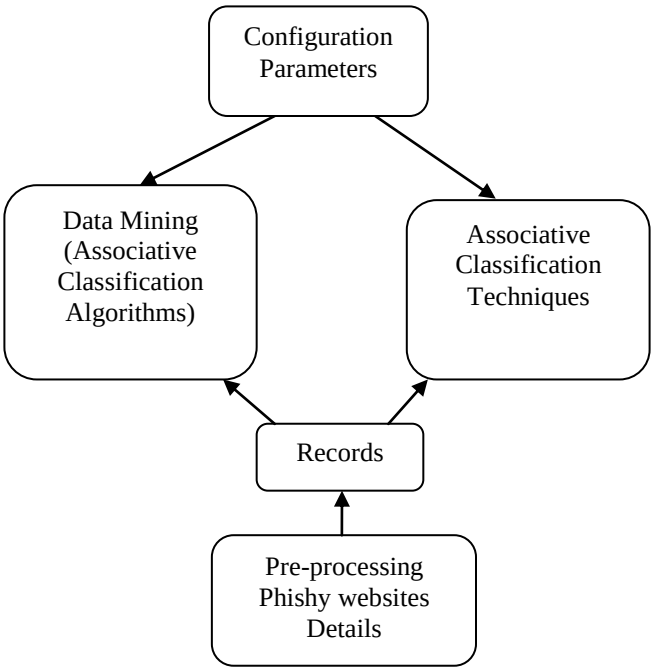

Fig 1: AC Model for Detecting Phishing Websites

In fig 1, a number of different existing data mining association and classification techniques including JRip [8], PART [8], PRISM [9] and C4.5 [10], CBA [11], MCAR [12] algorithms to learn and to compare the relationships of the different phishing classification features and rules.

III. RELATED WORK

Phishing website is a critical problem these days, due to its huge impact on the financial and on-line vending sectors and however, preventing such attacks is a significant step towards defending against website phishing attacks. There are several favorable approaches [7] to this problem and a broad collection of related works. There are various techniques which defend against phishing. Some techniques give e-mail level protection and some provide security toolbars embedded with anti-phishing tools.

Phishing [4] is a growing problem on the internet today for both consumers and businesses. One of the most common approaches for an attacker is to create a similar website in

order to capture personal information from consumers. A malicious website may appear identical to an online bank or other financial institution in order to capture passwords, social security numbers, account numbers, and other confidential information. A victim may not identify the malicious sites until once the confidential information has been leaked.

Some of the approaches for phishing detection are:

1. Email-level approach

This approach intends to amend the phishing attacks at the email level. The main concept is that when a spoofed email is not received by its victims, they cannot fall for the scam. Filters and content analysis techniques are often used to detect phishing emails before they can be delivered to users. For instance, by using training filters (e.g., Bayesian filters), an enormous number of phishing emails can be thwarted. In order to prevent spoofing of sender information in an email message, Microsoft and Yahoo have defined email authentication protocols (Sender ID and Domain Keys) that can be used to verify the credibility of a received email.
If widely used, these solutions could help to prevent spam emails and, as a result, decrease the number of email based phishing attacks.

2. Browser-integrated tool approach

These tools detect phishing by comparing the web page link in the address bar with the list of malicious site URLs mentioned in a blacklist. For example, the address bar turns red in Microsoft Internet Explorer (IE) 7 when a malicious page loads. Well-known, academic, browser-integrated solutions to slacken phishing attacks are SpoofGuard and PwdHash, SpoofGuard works by analyzing for phishing symptoms such as obfuscated URLS in web pages. On the other hand, PwdHash generates domain-specific passwords that are rendered ineffectual when submitted to another domain.

3. Webpage content analysis

It analyzes a Web page's content, such as the HTML code, text, images, input fields, forms and hyperlinks. Earlier, such content based approaches proved effective in detecting phishing pages. But recently, phishers have started creating web pages with non-HTML components, such as Flash objects, images and Java applets. For instance, a phisher might design a fake page that consists entirely of images, even if the original page contains only text information. In this case, content-based anti-phishing tools cannot analyze the suspicious webpage because its HTML code contains nothing but HTML <img/>elements.

4. Visual similarity based approach

Liu et al. short paper suggests that authors define metrics by analyzing and comparing legitimate and phishing web pages which can be used to detect a phishing page. The idea is to first disintegrate the web pages into pertinent blocks according to "visual cues." Then, based on the defined metrics, similarity between two web pages is determined. If the resemblance to the legitimate web page is above the predefined threshold, then the web page is considered as a phishing page.

In this section, the current anti-phishing approaches [2] are surveyed, and classified into two groups. These are: Blacklist/whitelist approaches and pattern recognition approaches.

Blacklist/White list Approaches

Ludl et al. [2] measured the effectiveness of two popular blacklist based approaches. These are: the blacklist preserved by Google and used by Firefox, and the Whitelist preserved by Microsoft and used by Internet Explorer. Their results show that Google was able to label 90% correctly, but Microsoft labels only 67%. Sharif et al. proposed a phishing blacklist website approach that avoids the problem of keeping the blacklist updated. Their proposed approach can be installed on the mail server to identify the set of URLs in an email, and the attacked company name. The authors have performed an experiment to compare the URLs collected from the email with that of the actual company took from Google search engine. The results show that their approach can score about 100% accuracy in detection phishing URLs with 9% of false positive. Though, the authors did not show how they can get the logo of the companies worldwide or how image comparison was performed. Furthermore, they did not show how to deal with URL address that is hidden by a proxy which limits the practicality of their study. Sheng et al. revealed that blacklists are updated at various speeds. They estimated that 47% - 83% of phishing URLs are added to blacklists 12 hours after performing experiment. Moreover, the authors found that zero hours protection delivered by major blacklist-based toolbars claims a true positive between 15% and 40%. So it is mandatory for a decent blacklist to be updated instantly.

The opposite term to blacklist is whitelist, which is a set of trusted websites, while all other websites are considered bad or untrusted. Chen and Guo [2] proposed an anti-phishing approach called Automated-Individual Whitelist (AIWL), based on an individual user's whitelist of known trusted sites. AIWL trace every login attempts performed by the users individually using a Naive Bayesian classifier. In case a repeated successful login for a specific website achieved, AIWL prompts the user to add the website to the whitelist. Users are warned once they submit their credentials to a website that does not exist in the whitelist. This technique assumes that users solely repeatedly submit credentials to legitimate sites, however all other sites are considered malicious.

Pattern Recognition Approaches

Abu-Nimeh et al. presented a study that compares the effectiveness of six machine learning approaches in detecting phishing emails. These are: Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NN). The authors collected a data set that consists of 1171 raw phishing emails and 1718 legitimate emails. Each email is represented by 43 features while effectiveness is measured by a weight error term which gives a higher weight to false negatives than false positives. This approach of measuring effectiveness is widely used for spam filter because the effect of considering a legitimate email as spam is worse than letting a spam email pass to the client mailbox. The authors applied a weighted error measure that considers a false positive 9 times more costly than a false negative.

Generally and according to Patil et al. [6] there are two types of classification problems, these are termed as single label and multi-label. In a single label classification, each training case in the input data is associated with only one class. In cases where the input data set contains just two class labels, the problem is called binary classification. However, if more than two classes are available, the problem is named multi-class classification.

The majority of existing AC mining algorithms use rules learnt from the training data set for constructing a single label classifier which in turn is utilized for predicting test data. Thus, there are limited numbers of research articles related to multilabel rules in AC. Hereunder, some of the authors shed light on two approaches and other techniques related to traditional multilabel classification in data mining.

Over the past decade, many researchers have investigated the problem of detecting phishing websites using data mining techniques. In this section, the author's shed light on both traditional data mining techniques and AC approaches.

Abdelhamid et al. [13] investigated the problem of website phishing using a new proposed multi-label classifier-based associative classification, MCAC. The main goal of the MCAC algorithm developed is to recognize attributes or features that distinguish phishing websites from legitimate ones. The table 1, results showed that the MCAC algorithm forecasted phishing websites better than traditional data mining algorithms.

Dadkhah et al. [14] developed a new method to forecast and detect phishing websites using classification algorithms based on the weight of web page features. The results showed that the proposed method produced a lower error rate than other data mining methods.

Abdelhamid [13] proposed an enhanced multi-label classifier-based associative classification algorithm, eMCAC. This generates rules associated with a set of classes from single-label datasets using the transaction ID list (Tid-list) vertical mining approach. The algorithm employs a novel classifier building method that reduces the number of generated rules.

Jabri and Ibrahim [15] proposed an enhanced PRISM algorithm for forecasting phishing websites. The experimental results revealed that the modified PRISM algorithm outperformed the original PRISM algorithm in terms of the number of rules, accuracy (87%), and lower error rate (0.1%).Alazaidah et al. [16] proposed a new multi-label classification algorithm based on correlations among labels, MLC-ACL. The MLC-ACL utilizes both problem transformation techniques and algorithm adaptation techniques. The proposed algorithm starts by converting a multi-label dataset into a single-label dataset using the least frequent label criteria, and then employs the PART machine learning classifier on the converted dataset. The output of the classifier is multi-label rules. In addition, MLC-ACL attempts to gain advantage from positive correlations among labels using the predictive Apriori algorithm. The MLC-ACL algorithm was investigated using two multi-label datasets named Emotions and Yeast. The experiments revealed that the MLC-ACL algorithm outperformed other machine learning algorithms in terms of three well-known evaluation measures (Hamming Loss, Harmonic Mean, and Accuracy).

Taware et al. [17] proposed a new MCAC that aims to recognize attributes that differentiate phishing websites from legitimate ones. The MCAC algorithm produced better results than other data mining algorithms with regard to accuracy.

Table 1: Summary of existing researches in the area of Phishing Detection using Data Mining

Authors	Contribution Summary	Result/ Weakness	Mechanism	Algorithm
Abdelhamid et al. [13]	Investigated the problem of website phishing using a new proposed multi-label classifier-based associative classification, MCAC. The main goal of the MCAC algorithm developed is to recognize attributes or features that distinguish phishing websites from legitimate ones.	The results showed that the MCAC algorithm forecasted phishing websites better than traditional data mining algorithms.	The experiments are conducted on a number classification data set related to website phishing to evaluate the eMCAC algorithm performance.	C4.5, PART and RIPPER
Dadkhah et al. [14]	Developed a new method to forecast and detect phishing websites using classification algorithms based on the weight of web page features.	The results showed that the proposed method produced a lower error rate than other data mining methods.	In this technique considering all possible parameters which are influential in detection of phishing attacks, in addition, compared to other similar techniques which use classification algorithms, it can detect journal phishing and it has a lower error rate.	Social engineering tool, C4.5, C and R Tree.
Alazaidah et al. [16]	Proposed a new multi-label classification algorithm based on correlations among labels, MLC-ACL. The MLC-ACL utilizes both problem transformation techniques and algorithm adaptation techniques.	The experiments revealed that the MLC-ACL algorithm outperformed other machine learning algorithms in terms of three well-known evaluation measures (Hamming Loss, Harmonic Mean, and Accuracy).	In this method they have taken two different application domain data sets; Biological, and Musical. The first dataset is called "Emotions" and it is concerned about songs according to the emotions they evoke. This data set contains six labels, with label cardinality (LC) equal to 1.869and label density (LD) equal to 0.311.	C4.5, PART
Jabri and Ibrahim [15]	Proposed an enhanced PRISM algorithm for forecasting phishing websites.	The experimental results revealed that the modified PRISM algorithm outperformed the original PRISM algorithm in terms of the number of rules, accuracy (87%), and lower error rate (0.1%).	Proposed a new phishing websites detection model based on PRISM algorithm, the prediction of phishing websites is essential, and this can be done using data mining classification algorithms, the classification system automatically classify phishing pages.	CBA, PART, C4.5, MCAC, JRip, MCAR, PRISM
Taware et al. [17]	Proposed a new MCAC that aims to recognize attributes that differentiate phishing websites from legitimate ones.	Proposed a new MCAC that aims to recognize attributes that differentiate phishing websites from legitimate ones.	The system goal is to detect phishy website by using MCAC algorithm. The MCAC algorithm generates rules further that rules are sorted by using sorting algorithm.	MCAC, C4.5, PART
Antonelli et al. [18]	Developed a new efficient AC algorithm using a fuzzy frequent pattern method.	The experiment results showed that the new fuzzy AC algorithm outperformed the well-known CMAR algorithm and generated accuracies similar to two recent AC algorithms, namely FARC-HD and D-MOFARC, on 17 real-world datasets.	Presented associative classifier based on a 459 fuzzy frequent pattern (AC-FFP) mining algorithm. AC-FFP consists 460 of the following three phases: Discretization, Fuzzy CAR Mining, and Pruning.	FP-Tree, CAR, C4.5
Isredza Rahmi et al.[19]	Using Feature Selection and Classification Scheme for Automating Phishing Email Detection	The feature selection used in this paper does not work on graphical form as some attacker bypass the content based approach using images.	When an email is sent, the message is routed from sender's server to the recipient's email server through MTA (Mail Transport Agent). MTA handles message transportation and acts as sorting area and mail carrier. This is where every email messages is stamped with email header information including message-id. This part of email header is not visible to most users but it is a useful indicator in determining phishing email.	Bayes Net algorithm, Simulated Annealing search algorithm

Antonelli et al. [18] developed a new efficient AC algorithm using a fuzzy frequent pattern method. The experiment results showed that the new fuzzy AC algorithm outperformed the well-known CMAR algorithm and generated accuracies similar to two recent AC algorithms,

namely FARC-HD and D-MOFARC, on 17 real-world datasets.

Isredza Rahmi A Hamid, Jemal Abawajy and Tai-hoon Kim [19] used hybrid feature selection method to detect phishing email. The main objective is to identify the behavior features in phishing email. This approach is based on the message provided in the message_id field. The message_id tags provided in the email header is used to verify the sender behavior. Using hybrid feature selection algorithm, 7 features are taken from the email. The author uses these features to mine the sender behavior to identify whether the email came from legitimate source or not.

REFERENCES

[1] Mitesh Dedakia, Khushali Mistry, "*Phishing Detection using Content Based Associative Classification Data Mining*", Journal of Engineering Computers and Applied Sciences, Vol.4, No.7, pp.1-11, 2015.

[2] Jaswant Meena, Ashish Mandloi , "*Classification of Data Mining Techniques for Weather Prediction*", International Journal of Scientific Research in Computer Science and Engineering, Vol.4, Issue.1, pp.21-24, 2016.

[3] Suzan Wedyan, Fadi Wedyan, "*An Associative Classification Data Mining Approach for Detecting Phishing Websites*", Journal of Emerging Trends in Computing and Information Science, Vol. 4, No. 12, pp.23-37, 2013.

[4] N.M. Shekokar, Chaitali Shah, Mrunal Mahajan, Shruti Rachh, "*An Ideal Approach for Detection and Prevention of Phishing Attacks*", Procedia Computer Science, Vol.4, Issue.9, pp.82-91, 2015.

[5] Rajendra Gupta, Piyush Shukla, "*Performance Analysis of Anti-Phishing Tools and Study of Classification Data Mining Algorithms for a Novel Anti-Phishing System*", I. J. Computer Network and Information Security,Vol.7, No.12, 70-77, 2015,.

[6] R.V. Patil, S.S. Sannakki, V.S. Rajpurohit, "A Survey on Classification of Liver Diseases using Image Processing and Data Mining Techniques", International Journal of Computer Sciences and Engineering, Vol.5, Issue.3, pp.29-34, 2017.

[7] M. Aburrous, M. A. Hossain, K. Dahal, F. Thabtah, "*Predicting Phishing Websites Using Classification Mining Techniques with Experimental Case Studies*", 2010 Seventh International Conference on Information Technology: New Generations, Las Vegas, pp. 176-181, 2010.

[8] I.H. Witten, E. Frank, "*Data Mining: Practical machine learning tools and techniques* (2nd Edition)", Morgan Kaufmann, San Francisco, pp.1-32, 2005.

[9] J. Cendrowska, "*PRISM: An algorithm for inducing modular rule*", International Journal of Man-Machine Studies, Vol.27, No.4, pp.349-370, 1987.

[10] J. R. Quinlan, "*Improved use of continuous attributes in c4.5*", Journal of Artificial Intelligence Research, Vol.4, Issue.3, pp.77-90, 1996.

[11] 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, pp.56-73, 1998.

[12] T. Fadi, C.Peter and Y. Peng, "*MCAR: Multi-class Classification based on Association Rule*", IEEE International Conference on Computer Systems and Applications, China, pp. 127-133, 2005.

[13] N. Abdelhamid, A. Ayesh, F. Thabtah, "*Phishing detection based Associative classification data mining*", Expert Syst. Appl., Vol.41, Issue.13, pp.5948-5959, 1987.

[14] M. Dadkhah, M. Jazi, V. Lyashenko, "*Prediction of phishing websites using classification algorithms based on weight of web pages characteristics"*, J. Math. Technol., Vol.5, Issue.2, pp.24-35, 2014.

[15] R. Jabri, B. Ibrahim, "*Phishing websites detection using data mining classification model*", Trans. Mach. Learn. Artif. Intell., Vol.3, Issue.4, pp.42-51, 2015.

[16] R. Alazaidah, F. Thabtah, Q. Al-Radaideh, "*A multi-label classification approach based on correlations among labels*", Int. J. Adv. Comput. Sci. Appl., Vol.6, Issue.2, pp.52–59, 2015.

[17] S. Taware, C. Ghorpade, P. Shah, N. Lonkar, *"Phish detect: detection of phishing websites based on associative classification*", Int. J.Adv. Res. Comput. Sci. Eng. Inf. Technol., Vol.4, Issue.3, pp.384-395, 2015.

[18] M. Antonelli, P. Ducange, F. Marcelloni, A. Segatori, "*A novel associative classification model based on a fuzzy frequent pattern mining algorithm"*, Expert Syst. Appl., Vol.42, Issue.4, pp.2086-2097, 2015.

[19] IRA. Hamid, Jemal Abawajy, Taihoon Kim, *"Using Feature Selection and Classification Scheme for Automating Phishing Email Detection"*, Studies in Informatics and Control, Vol.22, Issue.1, pp.61-70, 2013.

Authors Profile

Meenu Shukla,pursued Bachlors of engineering from Madhav Institute of Technology and Science,Gwalior (MP), India in 2014.She is currently pursuing her Master of Engineering from Madhav Institute of Technology and Science,Gwalior (MP),India

Dr. Sanjiv Sharma PhD, M.Tech(IT), B.E.(IT) is an Assistant Professor in the Department of Computer Science Engineering and Information Technology at Madhav Institute of Technology and Science Gwalior (MP), India. He received his PhD degree (Computer Science and Engineering) from Banasthali University, Jaipur (Raj.), India in 2014 and M.Tech (Information Technology) with honors from School Of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal(MP), India in 2007. His current research interests include Social Network Analysis, Data Mining, Network Security and Adhoc Network and Mobile Computing and their interdependency..

