IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-7 E-ISSN: 2347-2693

Performing Efficient Phishing Webpage Detection

Samanjeet Kaur[1], Sukhwinder Sharma[2]

CSE Department, BBSBEC, India,

www.ijcseonline.org

Received: Jun/21/2015 Revised: July/11/2015 Accepted: July/24/2015 Published: July/30/ 2015

***Abstract*—** Along with deployment of internet, saving financial and sensitive information becomes more inconvenient. One of the problems faced today is growing number of phishing websites. Phishing websites are fake webpage shaped and used by phishers to copy the web pages of legitimate websites which results in lack of faith in internet based services and causes financial loss to the internet users. So it has become crucial to search for useful solution applicable for such phishing websites. Therefore, establishing useful solution for mitigating phishing websites is essential to reduce the incident of being victimized by phishing attack. This research paper employs approach that uses fuzzy logic with classifiers like SVM, NMC and Gaussian. Fuzzy based detection system provides effective aid in detecting phishing websites. It successfully resulted in low false positive and high true positive for classifying phishing websites.

***Keywords*—** Phishing;SVM;Gaussian; Fuzzy Logic;Feature collection

I. INTRODUCTION

Phishing attack causes large scale security risk to the online community and for those who deal with the sensitive information for the reason that phishers makes identical copies of the website to direct the users to forged site that steals the information. Even if the web users are conscious of these types of attacks, then also lot of users become victimized under this attack of phishing. Only professionals or experts can recognize these types of fraudulent websites. Not all the web users are expert in recognizing them immediately; therefore web user becomes victim as a result of providing personal details to the attacker. Phishing is developing constantly as it is easy for attackers to make replica of entire website using HTML source code. By doing little changes in the source code of the website, it becomes easy to befool the victim by directing them to phishing websites. Moreover phishers make use of techniques that attract the web users, they use Greetings which attract web customers to verify their account right now without any delay or to update them otherwise their account will be terminated. According to the report of The Economic Times By 2015 the cyber crime in the country might double to 3 Lakh as a result which would cause serious threat to national security and economy. The growing use of tablets and smart phones for financial transactions, online banking and other have increased risks. India is highly preferred among mostly hackers, cybercriminals and other malicious users who make use of the internet to commit crimes like phishing, Identity theft, spamming and other types of fraud. Online banking accounts or cloning of ATM/debit cards are common occurrences of Phishing attacks. Greatest number of attackers belong to the 18-30 age group, added the report. With growing use of information technology (IT) enabled services for example online business, e-governance, protection of personal, electronic transactions and sensitive information have assumed chief importance. As per the findings, entire number of cyber crimes registered during 2011, 2012, 2013 and 2014 stood at 13,301, 22,060, 71,780 and 1,49,254 respectively.. The study said India ranks third after Japan and US in the list of countries mainly affected by online banking malware during 2014. On May 6, 2015 CYREN published its Q1 2015 Cyber Threat Report. In the report, CYREN security analysts note a steep rise in phishing URLs, tracking 3.86 million at the end of March compared to 2.55 million at the start of the year – a 51% increase [2, 3].

Various methods are being implemented at present for identifying phishing websites. Aburous et al [1] proposed an approach using fuzzy data mining for intelligent detection of phishing website. E-banking phishing detection is being performed on URL and domain identity, Page style and content, Web address bar, Security and encryption, Social human factor and Source code and JavaScript. Basnet et al. [4] adopted machine learning approach for detecting phishing. SVM and neural network are used for predicting phishing emails. It classifies phishing email by employing structural features in email and by using machine learning algorithms. Markopoulou [5] used lexical features to predict the phishing website. Lexical Features accuracy is compared with accuracy of hand selected automatically selected features. Algorithms like SVM, Online perceptron etc are used for prediction. In the work by Santhana lakshmi [6] machine learning algorithms have been used. Third party services like search engine, blacklist are used mainly for predicting phishing websites. Algorithms like MLP

(multi layer perceptron), NB (naïve bayes) and DT (decision tree) are used. Processes of feature and identity extraction are used and numbers of experiments are carried to identify performance of models. Mingxing et al. [16] Invented efficient phishing webpage detector in which 12 features are being used to determine whether webpage is phishing or legitimate.

II. MODEL FOR PHISHING WEBSITE DETECTION

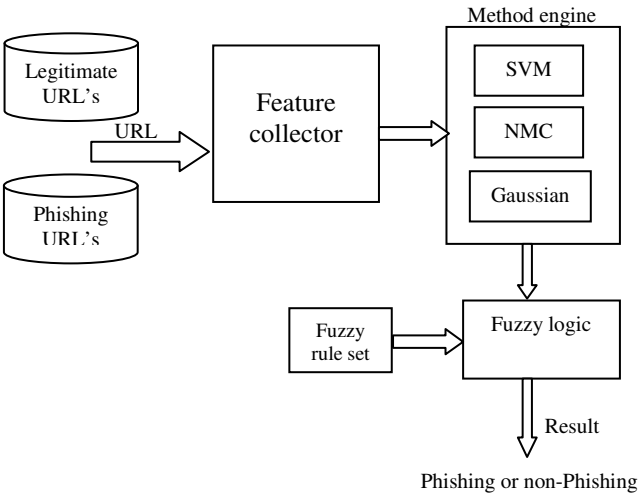

Fig 1: Flow diagram of proposed phishing detection method

A hybrid system is proposed which combines and integrates fuzzy logic with different classifiers. The proposed methodology uses classifiers like SVM, Nearest mean and Gaussian with fuzzy logic. Features will be collected to differentiate legitimate and phishing websites. Features will be obtained from areas like URL address of website. Features collected will be input to the method engine which contains SVM, NMC and Gaussian from where the rules will be generated that will be input to the fuzzy logic. In this research, by analyzing the existing methods of phishing detection and understanding the limitations, an efficient method will be proposed that can reduce the false positive (False positive rate measures the percentage of legitimate pages which are falsely labeled phishing) and can improve true positive (number of phishing pages which are labeled as phishing).

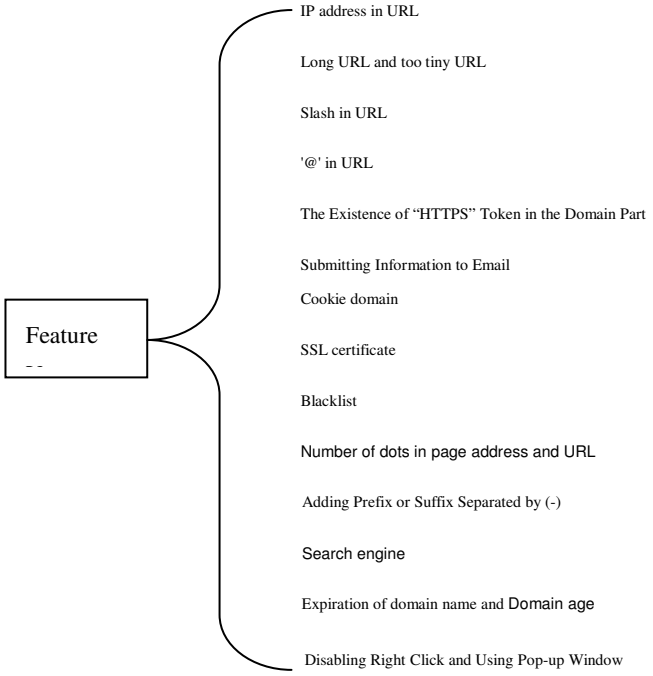

Fig2: The Feature Vector of the Proposed Model

A. *Feature vector*

It refer to the following information items in target website, including [7, 8]:

- *IP address in URL*

If the URL of a target website enclose an IP address as an alternative to domain name. Then there are more chances of website to be considered as phishing.

- *Slash in URL*

There should not be more number of slashes if the numbers of slashes are more than five then the URL is considered as phishing.

- *Long URL*

Web page with short URL is more trustworthy than that the page with suspicious long URL. To keep in view accuracy of study, length of URLs in the dataset is calculated and an average URL length is produced. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing [8].

- *'@' in URL*

Presence of '@' in the page address indicates that website is phishing since its presence indicates that text before '@' is comment.

- *Submitting Information to Email*

Web form permits user to give his personal information that is directed to a server for processing. A phisher may redirect the user's information to his personal email. To that stop, a server-side script language may be used such as "mail()" function in PHP. Client-side function "mailto:" might be used for this purpose. Thus presence of "mail()" or "mailto:" indicates site as phishing.

- *The Existence of "HTTPS" Token in the Domain Part of the URL*

The phisher may insert the "HTTPS" to the domain part of a URL in order to mislead users.

For example: http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

- *Cookie*

Web Cookie is usually used by the legitimate or original website to transmit the state information to the user's browser as well as by browser to give back the state information to original site.

- *SSL certificate*

SSL refers to secure socket layer. It creates secure connection between server's and user's browser permitting the private information to be conveyed without the trouble of eavesdropping. Legitimate website will have SSL certificate. However phishing websites do not have SSL certificate. If the certificate doesn't exist then site is considered as phishing otherwise legitimate.

- *Blacklist*

It is third party service which contains record of suspected websites. URL of the page is checked alongside the blacklist. If the URL of the page is present in blacklist, it is considered as phishing otherwise legitimate.

- *Number of dots in page address*

If number of dots in page address and number of dots in URL in source code are more than five it is considered as phishing website and all the dots are checked.

- *Search engine*

If the site is legitimate and URL of the page is assigned to search engine then first 5 results generated will be regarding concerned website. If the URL is fake then no results will be generated regarding concerned website.

- *Expiration of domain name*

Expiration of the domain name, symbolize that number of days left before a domain name expires, as sooner as domain name will expire the more likely that it is a phishing website.

- *Domain age*

It is usually represented by the number of days ever since the domain name was registered. The earlier the date that a domain name was registered to the current date, the more probable that it is a phishing website.

- Adding Prefix or Suffix Separated by (-) to the Domain

The (-) dash symbol is not often used in legitimate URLs. Phishers add prefixes or suffixes separated by (-) to the domain name. So that users believe that they are dealing with a legitimate webpage.

For example: http://www.Confirme-paypal.com/.

- *Using URL Shortening Services "TinyURL"*

URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage.

- *Disabling right click*

Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. This feature is treated exactly as "Using onMouseOver to hide the Link". Nonetheless, for this feature, we will search for event "event.button==2" in the webpage source code and check if the right click is disabled.

- *Using Pop-up Window*

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

B. Classifiers used

In this study, we have used three different classifiers for the detection of phishing websites. Those include SVM, Nearest Mean and Gaussian to build better phishing detection model. In this section, we will briefly introduce these.

Support vector machine (SVM) [9] a well known data classification technique, to classify webpage features, original SVM algorithm was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. SVM might find a separating hyperplane in the m-dimension space, which disconnect X into two classes such that vectors on one side of hyperplane have label 1 and vectors on the other side have label -1. Since a webpage is only considered as a legitimate or a phishing, it is naturally a binary classification problem. SVM would generate output in two classes. Nearest Mean (NMC) is simplest classifier introduced by Fukunaga (1990) as a classifier with lower complexity. It is also called (Euclidean distance) classifier. NMC provides good performance for small sample size problem. Gaussian classifier assumes that the observations are generated by a random process that has normal distribution. Density function of a normal distribution is defined by mean vector, and covariance matrix [10].

III. EVALUATION OF THE PROPOSED MODEL

A. *Data Collection*

Data set is structured from two websites, ‘‘phishTank” from the phishtank.com which is one of the most important phishing-report collector, for a total of almost 100 phishing websites are taken from phishTank[11]. The PhishTank database collects the URL for the website that are suspected as phishing are being reported, the time of that report, and sometimes detail such as the screenshots of the website and is publicly available. Moreover, the Anti-Phishing Working Group (APWG) [12] keeps a ‘‘Phishing Archive” unfolding phishing attacks. In addition, legitimate websites were collected from yahoo directory and starting point directory. Both directories contain addresses of legitimate websites for different types of services.

B. *Performance metrics*

Two metrics used to calculate the performance, which are True positive (TP) rate and false positive (FP) rate.

- *True positive*: It measures the percentage of phishing pages which are labeled as phishing. The higher TP value represents the better detector and it is computed by

  TPR = TP/(TP+FN)

- *False positive:* It measures the percentage of legitimate pages which are falsely labeled as phishing. The lower FP value the better detector and it is computed by

  FPR=FP/(FP+TN)

IV. RESULTS AND DISCUSSION

The purpose of this following experiment is to determine what percentage of phishing and legitimate URLs would be detected. The performance of the proposed system is calculated. For implementation, MATLAB® tool has been used. Table1 represent the results in form of TP (True Positive), FP (False positive).

Technique used	True positive	False positive
Fuzzy Logic combination with classifiers	98.9%	1.03%

Table 1: Results for True positive and False positive

In methodology classifiers are integrated with fuzzy logic to improve the overall result. Thus, our method proved effective as result of combination gives valuable results. True positive rate comes out to be 98.9% which is effective as true positive should be higher than false positive that comes out to be 1.03%.

.

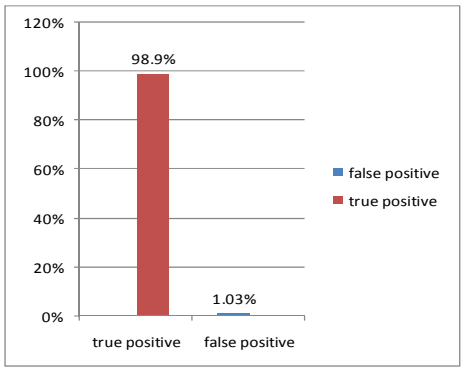

Fig 3: Represents overall methodology of integrating fuzzy logic with classifiers

V. CONCLUSION

In this paper a phishing detection approach is being proposed that classifies the webpage safety by examining the webpage address, phishing characteristics are extracted to estimate the security of the website. Finally True positive and false positive rate is calculated. Combined outcome is calculated by integrating classifiers like SVM, NMC, Gaussian with fuzzy logic it is found that true positive rate comes out to be 98.9 which is higher than false positive is 1.03 which indicates that fuzzy logic is powerful tool in decision making and combining it with classifiers gives more valuable results.

REFERENCES

[1]. Aburrous, M., Hossain, M.A.; Thabatah, F.; and Dahal, K. (2008): Intelligent phishing website detection system using fuzzy techniques. Information and Communication Technologies: From Theory to Applications. ICTTA. April 7-11, **2008**, pp.1-6.

[2]. Cyren report<http://ir.cyren.com/releasedetail.cfm?releaseid=911178> accessed june **2015**.

[3]. The Economic Times news <http://economictimes.indiatimes.com/topic/phishing> accessed july **2015.**

[4]. Andrew H.Sung, Ram Basenet, Srinivas Mukkamala.(**2008**) "Detection of Phishing Attacks: A machine Learning Approach". InB. Parsad, editor, Springer ,Soft Computing Applications in Industry, Studies in Fuzziness and Soft Computing, volume 226, pages 373–383.

[5]. Anh Le,Athina Markopoulou,Michalis Faloutsos:PhishDef: URL Names Say it All.

[6]. Santhana Lakshmi V, Vijaya MS, (**2012**) : Efficient prediction of phishing websites using supervised learning algorithms, Procedia Engineering 30 pp.798 – 805.

[7]. Jiang, Hansi; Zhang, Dongsong; and Yan, Zhijun (**2013**):A Classification Model for Detection of Chinese Phishing E-Business Websites, PACIS Proceedings, Paper 152.

[8]. Rami M. Mohammad, Fadi Thabtah.(**2012**): An assessment of features related to phishing websites using an automated technique. Internet Technology and Secured Transactions, pp-492-497.

[9]. Adi Sutanto, Jui-Lin Lai, Muhammad Khurram Khan, Mingxing He, Pingzhi, Rong-Jian Chen,, Ray-Shine Run, Shi-Jinn Horng,(**2011**):An efficient phishing webpage detector. Expert Systems with Applications,38, pp.12018–12027.

[10]. Comparison of generalization properties of statistical and neural classifiers available at< http://www.mif.vu.lt/~valdas/PhD/TEXT/CHAPTER4.pdf>.

[11]. PhishTank. [Online]. Available from: http://www.phishtank.com/.

[12]. Antiphishing Working Group, Available from: www.antiphishing.org/reports/apwg-reports.

[13]. Zhang, Y., Hong, J., and Cranor, L. (**2007**). CANTINA: A Content-Based Approach to Detecting Phishing Websites. Proceedings of the 16th International Conference on World Wide Web. pp. 639-648.

[14]. Rashmi Gupta, Nivit Gill. (**2013**): Token Based Security for Prevention of Phishing Attack at Client Side. International Journal of Emerging Technologies in Computational and Applied Sciences, pp.184-189.

[15]. Ram B. Basnet, Andrew H. Sung, Quingzhong Liu.(**2014**): Learning to detect phishing urls. IJRET: International Journal of Research in Engineering and Technology, pp.11-14.

[16]. Mingxing He, Horng SJ, Fan P, Khan MK, Run RS, Lai JL, Chen RJ.(**2011**) "An efficient phishing webpage detector. Expert Systems with Applications", An International Journal. pp. 12018-12027.

