ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

A Comparative Study on Techniques BVM with SVM For Detecting Phishing Pages

Mr.K.Selvan[1],
Research Scholar,
PG and Research Department of Computer Science,
JJ College of Arts and Science(Autonomous),
Pudukkottai.

Dr.M.Vanitha[2]
Assistant Professor,
Department of Computer Science and Engineering,
Alagappa University,
Karaikudi

Abstract

Phishing is a serious s threat to internet security that occurs when a malicious website impersonates a legitimate one in order to acquire sensitive information such as User name, Password, Bank account number and secret pin code., This lead internet users and financial company peoples distrusted. The two most technical methods in fighting phishing attacks are the Blacklist and the heuristic based. The Majority of blacklist methods deal with phishing problem are based on Support Vector Machine(SVM). In this paper we present a comparative study on detecting phishing pages using BVM and SVM . Experimental results using real phishing and legitimate pages. Show that BVM detects fraud pages with higher accuracy. We also want to identify ultimate set of web page features .

Key Words : Phishing, Machine learning, Support Vector Machine.

1. Introduction

Phishing, just as its name implies is a way of internet fraud attacks, which send a mass of fraudulent E-mails and post forged web sites. After diddling the victims' trusts, the personal and private information, such as bank accounts, usernames, and passwords appear to be accessed through phishing scams. In modern society, as sales channels of various industries rely strongly on the Inter- net, phishing websites are being developed in a variety of forms and it is likely to disrupt every field it touches. An APWG (Anti-phishing Working Group) report shows that there were around 60,000 phishing websites appeared in email client or the websites were visited regularly by netizens [1]. Phishing scams have now overtaken the malicious URL injection to become the primary threat of web surfing security.

In previous studies of phishing detection, there were many automatic phishing detection methods that could be used to identify phishing websites accurately. A new approach is proposed to detect phishing websites through prioritizing common words by search engines, which estimates domain name and content relevance to check if the web page is legal, and the recognition rates of this method can reach 97% [2]. Kirda et al. [3] proposed the use of browser plug-in to analyze whether the user submission information is sensitive information in real time, and then set pre warning to users. Meanwhile, many anti phishing tools in the form of browser plug-in emerged in the Internet, such as phishing websites filter in Internet Explorer, Firefox browser from Google, Spoof Guard developed by Stanford University [4–6], and these tools are based on blacklist, white list, and heuristic analysis. When a visited website is judged to a phishing website, the tools will pop-up prompt box to warn users. Weiwei et al. [7] proposed the use of clustering ensemble method to aggregate multiple cluster algorithms, which can develop an automatic categorization system to automatically classify phishing websites. Dhamija et al. [8] proposed dynamic security skins method that is based on server schemes. Topkara et al. proposed a novel scheme 'ViWiD', which is an integrity check mechanism based on visible watermarking of logo images and its implementation for mitigating phishing attacks [9]. Ying and Xuhua [10] analyzed the properties of web page based on structural DOM model, and utilized support vector machine (SVM) in detecting phishing page, but this method has some limitation when dealing with image. Consider the spread of phishing website Abu- Nimeh et al. [11] proposed a feature extraction method that mainly extract the feature of the emails with phishing web pages, then evaluated the effects of six machine learning algorithms in terms of classification. Although this method improves the detection precision in detecting phishing web sites, it uses a single web page to extract features, and therefore, it is easy to be deceived by phishing website designers. Due to new phishing websites emerge in an endless stream, and they have the characteristics of cost-effective and short-lived, so the phishing detection methods need to have strong real-time performance and intelligence [12]. Tsang et al.

[13] proposed a method for detecting phishing pages by searching similar webpages through comparing the webpages by matching HTML source codes as well as computing the cosine similarity of detecting phishing pages. Most of the previous phishing webpage recognition techniques are aimed at single web page, the process of feature extraction depends on fewer and single page features, such as textual features, image features. These schemes are web-facing detection methods that cannot carry out comprehensive analyses on website. So, in

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

addition to reducing the detecting efficiency, that are unable to achieve the goal of real-time detecting phishing website. SVM is a traditional classification algorithm, as it showed to be higher in classification accuracy, but the SVM algorithm delivers slower training speed when coping with large training sets, and that becomes a drawback in practical applications [14].

For the above reasons, in this paper, we based on the differences between the phishing websites and the imitated target websites, from the perspective of the web topology structure. We use the BVM algorithm to detect and classify phishing site. In order to quantify the topological features, we first extract 12 statistic indices of web- sites as the topological features to complement the disadvantage of the detection based on single webpage features. Second, we have BVM instead of SVM, which then utilize BVM classifier to classify the features vectors, which are composed of topological features. As an independent scheme, the experimental results show that our method is efficient at detecting phishing sites, moreover, it can achieve higher accuracy of phishing detection and have faster training speed.

The rest of the paper is organized as follows: Section 2 describes in detail the theoretical basis of BVM classifier. Section 3 intro- duces the process of getting the raw data crawled by web crawler, obtaining the required features and forming the features vectors correspond to websites. Section 4 describes the datasets used in our experiments, and illustrates the experimental design and the performance of the proposed method. Conclusion and orientations for future works are discussed in section 5.

2. SVM classification via minimum enclosing ball

In recent years, many machine learning algorithms are devel oped for solving classification problems and have been widely applied for detection of phishing websites [15]. To construct high performance phishing detection model, there are two aspects of work need to be done: work need to be done: firstly, getting a good performance classifier and obtaining high precision in classifying normal data and anomaly data; secondly, reducing the training time without impacting the classification accuracy. The SVM performed good capability in detecting phishing website. In this paper, emphasis is given on the BVM algorithm, which is firstly applied the BVM algorithm to the phishing detection.

3. Phishing website detection based on BVM

Studies have found that the topologies of large websites are more complex than that of small websites, and the topologies of the phishing websites are surprisingly simple in general. Enlightened by the differences among the website topological structures, we utilize web crawler technology to fetch web pages [18]. To extract the topological features more efficiently, we construct the DOM tree in the website parsing process, extract the topological features by traversing the DOM tree. After that, the feature vectors of phishing websites are detected by BVM.

3.1. Analysis of website topology structure based on DOM tree

In order to quantify the website topology features, we select 12 kinds of web statistic indicators as the topological features of web- site, so the usage of quantitative indicators makes it easy to realize the standardization of feature vectors.

According to the characteristics of the website itself, each website can be transformed into a DOM tree [19]. Take a website and the topological features of the website can be obtained by the analysis of the DOM tree. In this paper, $F = /F_1, F_2, \ldots, F_8 /$ is defined as the feature vectors of websites. F_i is used to present its corresponding feature, which are described in detail as follows:

F_1 : IP Address. Using an IP address in the domain name of the URL is an indicator.

F_2:. Long URL. Phishers hide the suspicious part of the URL to redirect informations submitted by users.

F_3 : Adding prefix and suffix .Phishers try to scam users by reshaping the suspicious URL.

F_4 : Sub Domains. Another technique used by phishers to scam users is by adding a sub domain to the URL.

F5: SSL Final. The existence of HTTP Protocol every time sensitive information is being transferred is being transferred reflects that the user is certainly connected with an legitimate web page.

F6: DNS Record. An empty of missing DNS record of a website is classified as phishing.

F_7 : Using Pop up window. Usually authenticated web pages do not ask users to submit their credentials via a pop up window.

F_8 :Age of domain. Websites that have an online presence of less than 1 month can be considered.

3.2. Topological feature vectors extraction of websites

Web crawler technology is usually used in the search field. It can traverse webpage resources in the Internet as far as possible. However, it is not necessary to apply the ordinary web crawler tech- nology in phishing website detection. In this paper, we improve the ordinary web crawler workflow and extract the topological features by finitely traversing the specific websites.

As shown in Fig. 2, the main procedure of fetching the topological features by web crawler is given briefly as follows:

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

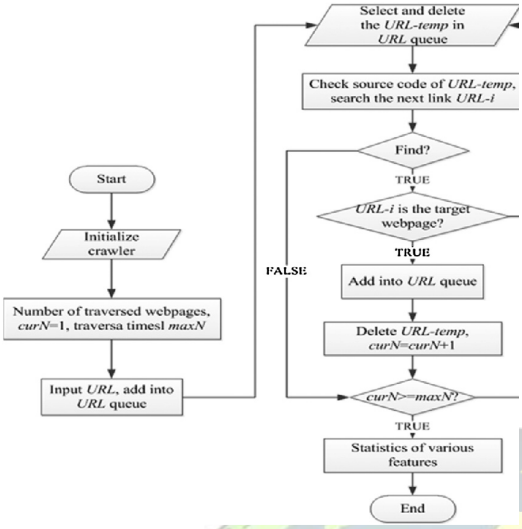


Fig. 2. The flowchart of web page feature extraction.

Step 1. Initialize the web crawler, set the number of traversed webpages curN = 1, the upper limit on the number of the traversed webpage within a website maxN;

Step 2. Input the URL of a page within in the website which need to be detected, record the website address which contains the URL, and then add the URL to the URL queue that have not been crawled;

Step 3. Remove a URL-temp from the URL queue, then delete the URL-temp from the queue, and set curN = curN + 1;

Step 4. Utilize web crawler to establish the http connection, get the URL-temp page source code;

Step 5. Use regular expression to search the URL-temp page, find the next link URL-i; If URL-i can be found, jump to the step 6. Otherwise, jump to the step 8;

Step 6. Determine whether URL-i is the Webpage of the currently detected website. If so, jump to the step 7. Otherwise, return to the step 5;

Step 7. Add URL-i into URL queue which have not been crawled, then return to the step 5;

Step 8. If curN ≥ maxN, the algorithm ends; Otherwise, return to the step 3;

Step 9. Analyze the source code of each webpage, get the topo- logical features.

In this paper, we can automatically collect the website topologi- cal features by utilizing the improved web crawler technology, take five phishing websites and five legitimate websites as an example, the topological features are respectively listed in table1.

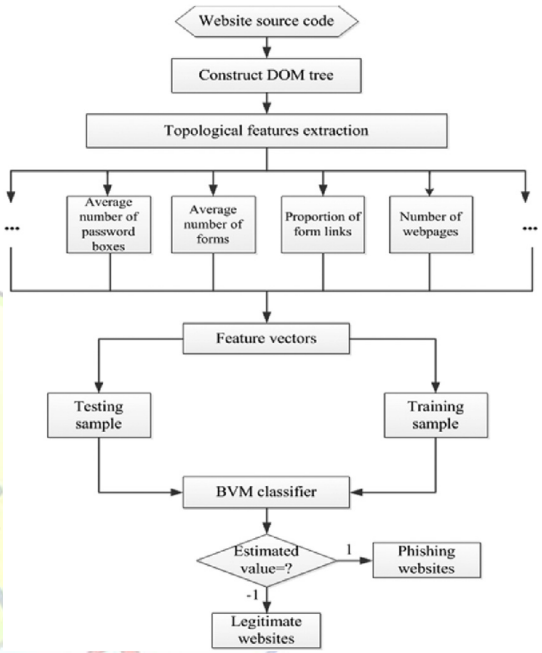


Fig 3 Architecture of the phishing detection based on BVM

3.3. BVM based phishing website classification

Support vector machine algorithm has become popular in phishing detection. However, as the number of training samples increase, the classification efficiency of the SVM for large-scale samples problem decline. Moreover, there are always other problems, such as higher memory usage and slower speeds. Compared with SVM, BVM has higher classification accuracy and lower training time in handling large-scale samples problem.

In order to obtain the minimum enclosing ball of one dataset, firstly, solve the enclosing ball of the dataset [20]; Secondly, get the minimum enclosing ball and SVM parameters. We construct BVM classifier and classify the website topological feature vectors after- wards. As shown in Fig. 3, the main procedure of phishing website classification based on BVM can be demonstrated clearly as follows:

Step 1. Obtain initial training samples through web crawler. Crawling the data of m legitimate websites by web crawler, where
m ≥ 200, and extracting the features to organize into a sample data;

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Then crawling the data of n phishing websites by web crawler, where n ≥ 200, and extracting the features to organize into a sample data;

Step 2. Label training samples. Set one binary response vari- able, which indicates that webpage is phishing (label with 1) or legitimate (label with -1).

Step 3. Extract the website topological features. Select 12 web- site topological features according to the obtained data crawled by the web crawler.

Step 4. Construct BVM classifier. In the process of solving theMEB of the datasets S, firstly, solve the EB (enclosing ball), that is to say, find a ball B(c, r), which contains all data points when r ≥ R*.

The (1 + ε) approximation algorithm for solving the EB are shown as shown Step4.1–Step4.3.

Step 4.1. Initialize c_0 = (x_0), if (x) is not in the outside of $B(c_t, (1+\varepsilon)r)$, the algorithm ends; Otherwise, select (x_1) as the arbitrary point in the outside of $B(c_t, (1+\varepsilon)r)$.

Step 5. Solve the minimum enclosing ball of the dataset S. Firstly, according to Eq. (9), let r = $\tilde{k}$; Secondly, according to Eq. (13), the solution of EB(S, $\tilde{k}$) is c = (ω , b, C), then the solution of the SVM can be solved as c = (ω , b, $\sqrt{C}$). Lastly, the BVM based classifier has been constructed.

Step 6. Detect the testing samples by BVM classifier, and output the results.

4. Experimental results and analysis

In order to evaluate the effectiveness and applicability of the proposed scheme, the experimental dataset, including phishing pages collected from phishtank.com and legitimate pages collected from the well-known classified websites. In this paper, we collect multiple websites and top-level domains, such as .com, .net, .org, .jp, and .uk. In the experiments, we first utilize the web crawler to collect 400 phishing websites and 400 legitimate websites, then extract 12 topological features from these website. We calculate the means and variances of the 12 topological features. Table 2 shows that the mean values and variances of phishing topological features are obviously distinct from those of legitimate topological features.

We use four standard measures, such as precision, false positive rate (FNR), Recall, F value to evaluate the proposed method.

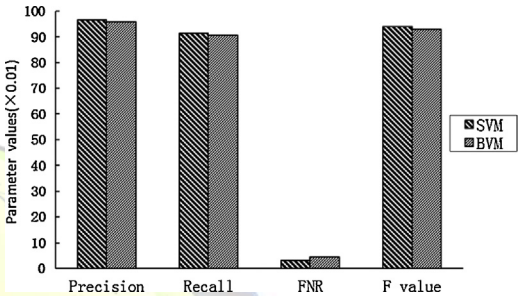
$$\text{Precision} = |TP| / (|TP| + |FP|)$$

$$\text{FNR} = |FN| / \{ |TP| + |FP|)$$

$$\text{Recall} = |TP| / (|TP| + |FN|)$$

$$F = (|\text{Precision}| * |\text{Recall}| * 2) / |\text{Precision}| + |\text{Recall}|$$

where TP is the number of legitimate pages which correctly classified as legitimate, TN is the number of phishing pages which correctly classified as phishing, FN is the number of legitimate pages, which misclassified as phishing, FN is the number of phish- ing pages which misclassified as legitimate, and F value is used to comprehensively reflect the whole performance of algorithm.



we utilize 400 phishing pages and 400 legitimate pages to generate BVM classifier and BVM classifier, respectively.

5. Conclusion

In this paper, we did a comparative study that BVM identifies phishing web pages along with its victimized features set that most of the other methods lack. The high performance in phishing page detection is achieved is possible because the BVM method narrow down the possible phishing pages. Though BVM achieves high performance, when web page is the combination of text and images, images would be taken as new criteria and run time performance are the future work.

References

[1] Anti-Phishing Working Group (APWG), Phishing activity trends report for the month of June, 2011 http://www.anti phishing.org/.

[2] A Machine Learning Approach for Detection of Phished Websites Using Neural Networks K. Selvan, M. Vanitha International Journal of Recent Technology and Engineering (IJRTE) SN: 2277-3878, Volume-4 Issue-6, January 2016.

[3] E. Kirda, C. Kruegel, Protecting users against phishing attacks with AntiPhish, in: Computer Software and Applications Conference, 2005 COMPSCA 29nd Annual, IEEE(the Institute of Electrical and Electronics Engineers), Piscataway, NJ, 2005, pp. 517–524.

[4] Microsoft Anti-phishing white paper, 2005.

[5] F. Schneider, N. Provos, R. Moll, M. Chew, B.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Rakowski, Phishing protection design documentation, 2007.

[6] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, J.C. Mitchell, Client-side defense against web-based identity theft, in: Proceedings of Symposium on Network and Distributed System Security, ISOC(Internet Society), New York, NY, 2004.

[7] Z. Weiwei, Y. Yanfang, C. Yong, L. Tao, Ensemble clustering for internet security applications, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Trans. Syst., Man, Cybern. C, Appl. Rev. 42 (2012) 1784–1796.

[8] R. Dhamija, J.D. Tygar, The battle against phishing: dynamic security skins, in: Proceedings of Symposium on Usable Privacy and Security, ACM(Association for Computing Machinery), New York, NY, 2005, pp. 77–88.

[9] M. Topkara, A. Kamra, M.J. Atallah, C. Nita-Rotaru, ViWiD: visible watermarking based defense against phishing, Lect. Notes Comput. Sci. 3710 (2005) 470–483.

[10] P. Ying, D. Xuhua, Anomaly based web phishing page detection, in: Computer Security Applications Conference, 2006 ACSAC '06 22nd Annual, IEEE(the Insti- tute of Electrical and Electronics Engineers), Piscataway, NJ, 2006, pp. 381–392.

[11] S. Abu-Nimeh, D. Nappa, X.L. Wang, S. Nair, A distributed architecture for phish- ing detection using Bayesian additive regression trees, in: eCrime Researchers Summit(eCrime), IEEE(the Institute of Electrical and Electronics Engineers), Piscataway, NJ, 2008, pp. 1–10.

[12] S. Marchal, J. Francois, R. State, T. Engel, Phish Strom: detecting phishing with streaming analytics, IEEE Trans. Netw. Service Manag. (2014), http:// dx.doi.org /10.1109/ TNSM.2014.2377295.

[13] W. Tsang, J.T. Kwok, P.-M. Cheung, Core vector machines: fast SVM training on very large sets, J. Mach. Learn. Res. (2005) 363–392.

[14] S. Roopak, T. Thomas, A novel phishing page detection mechanism using HTML source code comparison and cosine similarity, in: 2014 International Conference on Advances in Computing and Communications(ICACC), IEEE(the Institute of Electrical and Electronics Engineers), Piscataway, NJ, 2014, pp. 167–170.

[15] J. James, L. Sandhya, C. Thomas, Detection of phishing URLs using machine learning techniques, in: 2013 International Conference on Control Commu- nication and Computing (ICCC), IEEE(the Institute of Electrical and Electronics Engineers), Piscataway, NJ, 2013, pp. 304–309.

[16] V.N. Vapnik, Statistical Learning Theory, John Wiley and Sons Incorporated, New York, 1998.

[17] W. Yongqing, Z. Yongkang, Z. Suiwu, Simpler minimum enclosing ball: fast approximate MEB algorithm for extensive kernel methods, in: Control and Decision Conference, 2008 CCDC Chinese, IEEE(the Institute of Electrical and Electronics Engineers), Piscataway, NJ, 2008, pp. 3576–3581.

[18] P.GuptaK. Johari, Implementation of web crawlerin, in: 2009 2th Interna- tional Conference on Emerging Trends in Engineering and Technology(ICETET), IEEE(the Institute of Electrical and Electronics Engineers), Piscataway, NJ, 2009, pp. 838–843.

[19] K. Jinbeom, Y. Jaeyoung, C. Joongmin, Repetition-based web page segmentation by detecting tag patterns for small-screen devices, IEEE T. Consum. Electr. 56 (2010) 980–986.

[20] C. Fu-lai, D. Zhaohong, W. Shitong, From minimum enclosing ball to fast fuzzy inference system training on large datasets, IEEE Trans. Fuzzy Syst. 17 (2009) 173–184.

[21] W.T. Ivor, T.K. James, C. Pak-Ming, LIBCVM Toolkit-A Library for {Core/Ball} Vector Machines http:// www.cs. ust.hk /~ivor/cvm.html.

[22] C. Chih-Chung, L. Chih-Jen, LIBSVM-A Library for Support Vector Machines http://www. csie.ntu.edu. tw/~cjlin/libsvm/.

[23] L. Yongqian, S. Jie, Y. Yongping, L. Wei-Jen, Short-term wind-power prediction based on wavelet transform—support vector machine and statistic character- istics analysis, IEEE T. Ind. Appl. 48 (2012) 1136–1141.