A Genetic Algorithm Based Model for Chinese Phishing E-commerce Websites Detection

Zhijun Yan[1(✉)], Su Liu[1], Tianmei Wang[2], Baowen Sun[2], Hansi Jiang[1], and Hangzhou Yang[1]

[1] School of Management and Economics, Beijing Institute of Technology, Beijing, China
{yanzhijun,2120141523,2120111706,hangzhou}@bit.edu.cn
[2] School of Information, Central University of Finance and Economics, Beijing, China
{wangtianmei,sunbaowen}@cufe.edu.cn

Abstract. We propose a new Chinese phishing e-commerce websites detection model which integrates the URL features and web features of websites. Some unique features of Chinese e-Commerce websites are included and Sequential Minimal Optimization (SMO) algorithm is applied to identify the phishing e-commerce websites. At the same time, we adopt the genetic algorithm (GA) to optimize the detection model. The evaluation results show that the performance of SMO algorithm is better than the baseline model and GA improves the detection accuracy significantly.

Keywords: Chinese phishing website detection · E-commerce · Sequential minimal optimization · Genetic algorithm

1 Introduction

With the rapid development of Internet, e-commerce has gradually become an essential part of people's life. In 2013, the market transactions of China's online shopping exceeded 1.8 trillion RMB and annual growth rate was 39.4 % [1]. However, online shopping also results in a series of security problems, such as phishing attack. Phishing websites are usually spread by emails that look like coming from legitimate sources, and lure users to visit fraudulent websites through disguised URL. When the users disclose password and other account information in these phishing websites, their money will be transferred or stolen [2]. Between July 2011 and June 2012, 60 million Chinese online users became victims of phishing sites, and the cumulative loss was more than 30 billion RMB [3]. Therefore, it is important to develop an effective method to detect phishing websites and minimize consumers' financial loss.

Detecting phishing e-commerce websites is a challenging task. Phishing websites usually present professional webpages and provide similar sophisticated shopping process with real counterparts, making users difficult to distinguish real websites from fake ones [4]. Aiming to improve the accuracy of Chinese e-Commerce phishing websites detection, this paper proposes a new integrative approach by incorporating the unique features in Chinese e-commerce websites and applying the SMO and genetic algorithm to classify e-commerce phishing websites. Specifically, the proposed method defines the classification

F.F.-H. Nah and C.-H. Tan (Eds.): HCIBGO 2016, Part I, LNCS 9751, pp. 270–279, 2016.
DOI: 10.1007/978-3-319-39396-4_25

features from the view of URL features and web features, then the websites can be classified by the SMO algorithm, which is enhanced by the genetic algorithm. The proposed model neither needs expertise knowledge nor whitelist or blacklist, avoiding the maintenance work and increasing the reliability of classification system.

The rest of this paper is organized as follows. In Sect. 2, related works on the detection of phishing websites are introduced. Then, we propose a new Chinese e-commerce phishing websites detection model based on SMO and genetic algorithm. In Sect. 4, the experiment results are presented. Finally, we conclude our work.

2 Related Research

Existing phishing detection method can be roughly divided into four categories: URL blacklist based method, the visual similarity based method, the URL and text feature based method and the third-party search engine based method. We discuss the main result of these four types of research in the rest of this section.

URL blacklist based method is mainly based on a list of known phishing sites to identify phishing sites [5]. Some agencies or websites (such as PhishTank.com, Escrow-Fraud.com) maintains a blacklist, a collection of phishing sites that reported by Internet users around the world. If the URL of a target website is in the blacklist, it will be identified as a phishing site and blocked by application software. However, it is only used to prevent users from identified phishing sites and cannot detect new phishing sites. And you need to update the list constantly, which greatly increases the maintenance workload [6].

The visual similarity based method converts the detection of phishing sites to an image matching problem [7–9]. This kind of method assesses different website parts' similarity between the target website and the authentic website. If the similarity is higher the threshold value, the target website will be identified as phishing website. The visual similarity based method should divide the target website into different images, its detection performance lies in the development of web segmentation and image comparison algorithm.

URL and text feature based method identifies phishing sites according to the characteristics of URL and content characteristics of the target website [10–12]. By analyzing the sensitive characteristics of URL and text feature, it can distinguish the phishing website from the real website. The URL and text feature based method is the most common detection methods, but most of the existing detection models are generic method and do not include any context-related characteristics, which cannot have the best performance in specific domains.

The last kind of detection method is to search target URL information in third-party search engine, and then uses the collected information to make judgments [13, 14]. By comparing the search results with top and second level domain name of the target URL, it can identify the phishing websites. The big challenge faced by this method is that phishing site designer can optimize the search result of phishing sites, which makes this method invalid.

In summary, the current phishing website detection methods make great effort to detect phishing e-commerce websites using generic classification model, but they have various weakness. At the same time, as the fast growth of e-commerce, lots of small and

medium e-commerce companies emerge in China. Some of them vanish soon because of the highly competitive e-commerce environment. That also makes it infeasible to apply the previous methods to recognize and block phishing websites.

3 A Detection Model of Chinese Phishing E-commerce Websites

The proposed model incorporates the unique features of Chinese e-commerce websites, which are defined from the view of URL features and web features. Based on the defined feature vector, the SMO algorithm and genetic algorithm are applied to detect the phishing websites effectively. Different with the existing method, the proposed method does not rely on prior knowledge of real authentic websites, fits the e-commerce context of China, and has better classification accuracy.

3.1 The Phishing Website Feature Vector

By combining the prior website features used in literatures and new unique features of Chinese e-commerce websites, this study defines a feature vector for Chinese phishing e-commerce websites detection, which is divided into two parts: URL features and web features [15].

URL Features. URL features refer to a number of basic information extracted from the URL of a target website, which include the following sections:

IP-based URL: A phishing website URL usually uses IP address rather than a domain name, which can hide their real identification. For example, a phishing website may use http://121.73.1.108 to replace the URL of the official homepage of Jingdong.com, one of the largest B2C websites in China.
Presence of symbol '@': In the URL, the contents before the symbol '@' are the username and password for identity validation, and the content behind this symbol is the real address.
Presence of UNICODE characters: Phishing websites usually use UNICODE in their URL.
Number of dots ('.'): We can determine phishing sites by detecting whether the URL contains many '.' symbol.
Number of domain suffixes: The URL of a phishing website may contain many domain suffixes, such as.com,.cn,.org or other common Chinese domain name suffixes. For example, http://www.z.cn.1z.com.cn is a typical phishing site URL.
Age of domain name: The closer the date that a domain name was registered, the higher the possibility that it is a phishing website.
Expiration time of domain name: If the remaining valid date of a domain name is very short, it is likely to be a phishing website.

Consistence between DNS (Domain Name System) server address of domain name and URL: If the DNS server addresses of the domain name and URL are inconsistent, there may be a phishing site.

Registration status: By searching the MIIT website, we can find out whether the domain name of the target website is registered.

Registration subject: The website can be registered in MIIT by an individual or an enterprise. Considering the strict regulations on enterprise in China, the website registered by an individual has the higher probability to be a phishing one.

Registration site name: We can check whether the registered site name and actual site pointed by the URL are consistent.

WEB Features. Web features are obtained from website's source code through a web crawler. It includes the following sections:

Valid ICP (Internet Content Provider) certificate number: Real e-commerce websites will present ICP number at the bottom of the webpage, which is a unique identification issued by MIIT.

Number of void (null) links: Normally, the phishing website is likely to have more void links compared with authentic websites.

Number of out links: A phishing website tends to have more out links.

Valid e-commerce certificate information: In china, many authentic e-commerce websites receive certificates from industrial associations. They may post images of e-commerce certificates at the bottom of its website. Consumers can browse the detailed certificate information in industrial associations through these images.

3.2 Detection Algorithm

This study uses the machine learning algorithm SMO to detect Chinese phishing websites. SMO method is a simple algorithm [16]. It can quickly solve the Support Vector Machine (SVM) quadratic programming problems [17, 18]. For a binary classification problem with a dataset $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i is an input vector and y_i is a binary class label, a soft-margin support vector machine can be trained by solving a quadratic programming problem described as follows:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j$$
$$\text{subject to:} 0 \le \alpha_i \le C, \text{for } i = 1, 2, \ldots, n, \sum_{i=1}^{n} y_i \alpha_i = 0 \tag{1}$$

where C is an SVM hyperparameter (called penalty parameter) and $K(x_i, x_j)$ is the kernel function, both provided by the user; variables α_i and α_j are Lagrange multipliers. This optimization problem will be decomposed by SMO into a series of smallest possible sub-problems, and then solves them successively. Compared with other algorithms, the SMO method selects and solves a minimum optimization problem in each step. The major advantage of SMO approach is that the entire quadratic programming problem is broken into many small problems which completely avoided using the iterative algorithm. At the same time, its implementation doesn't require huge storage.

3.3 Model Parameter Optimization

Different kernel functions of SMO algorithm have a large impact on the classification accuracy [19]. The kernel function parameter r mainly affects the complexity degree of the sample's distribution in high-dimensional feature space, and the penalty parameter C is used to determine the level of confidence interval and experimental risk in a given feature space, and affect the SMO generalization capability.

In order to get the best algorithm performance, it is of vital importance to determine the appropriate combination of parameters for SMO algorithm. Genetic algorithm provides a general framework for solving complex system optimization problems. Based on the fitness function and genetic operators, the algorithm has the ability to reach the global optimization [20]. Prior literature showed that genetic algorithm has a good performance in parameters optimization [21]. However, it is rarely applied in phishing website detection. In this study, we used it to optimize the SMO parameters and identify phishing website more efficiently.

Chromosome Design. The first step of genetic algorithm is to design individual gene and its coding scheme. SMO algorithm is mainly related to three parameters: kernel function, kernel parameter r and penalty parameter C. In order to simplify the optimization and computation process, the chromosome is designed to be 31 genes. The first gene a1 represents the kernel function. Two widely adopted kernel functions are considered: the value 0 is for polynomial kernel function, the value 1 is for Gaussian kernel function. The penalty parameter C is represented by 15 genes, from a2 to a16, which describes that the range of penalty parameter C is from 0 to 327.68. Moreover, the kernel parameter r is described by 15 genes, from a17 to a31.

Fitness Function. The fitness function is objective function of the parameter optimization process. It is used to evaluate individuals' performance (fitness) in the search space. In this study, genetic algorithm is adopted to optimize the SMO parameters and provide a high degree of overall classification accuracy. Thus the overall accuracy of the classification model is defined as the fitness function.

Genetic Operators Design. The genetic algorithm has three basic operations: selection, crossover and mutation [22]. Selection makes sure that only some chromosomes of the population will be included in next generation. As the most common method, roulette wheel selection is used in this study. In roulette wheel method, the probability of an individual is included in the next generation is equal to the ratio of the fitness value of the individual and the entire population.

The crossover operation is conducted on the new population to improve the fitness of new population. It exchanges the gene at the same position on two different individuals (chromosomes), resulting in two new individuals. The single-point crossover method is applied, i.e., choosing an intersection point randomly and interchange the genes before and after the intersection. The default crossover rate is set as 0.75.

The mutation operator is helpful for finding the global optimal solution. It modifies the value of a random bit in the chromosome and improves the performance of the population resulting from crossover operation. In this study, the random selected

relevant bits should be mutated through change of every 0 bits to 1, and every 1 bits to 0. The default mutation rate is set as 0.2.

Parameter Optimization Process. At first, the initial population is randomly generated, which has 10 individuals. Then the SMO classification model is invoked and the fitness value of each individual is calculated. If the fitness value is low than 99 % and the iteration number doesn't reach 10,000 times, the selection, crossover and mutation operators will be applied in sequence. Thus the next generation is derived and SMO classification model will be called again. Iterate the above steps until the optimal parameters are gotten or the upper iteration time is reached.

4 Evaluation

4.1 Data Set

We have conducted an empirical evaluation of the proposed method by using the authentic and phishing e-commerce websites registered in third-party service platforms. Phishing e-commerce sites are from the online transaction security center (http://www.315online.com.cn) and Security Alliance (http://www.anquan.org), which validated and registered the phishing e-commerce websites complained by online consumers. Authentic e-commerce sites are collected from the online transaction security center. In order to optimize the training effect, the number of authentic and phishing websites are nearly same. Specifically, there are 1462 authentic e-commerce sites and 1416 phishing e-commerce sites.

A popular tool, called WebZIP, is used to download the source code of the collected e-commerce websites. Then the feature vector is extracted from the source code of online websites. We also used Weka (Waikato Environment for Knowledge Analysis), a widely adopted data mining tool, to train the proposed models.

4.2 Evaluation Metric

We use precision (P), recall (R), F-measure (F) and overall accuracy (O) as metrics to assess the effectiveness of the proposed detection model [23]. Specifically, precision is the percentage of correct detections. Recall measures the proportion of actual positives in the population being tested. The F-measure is a harmonic average of precision and recall, which represents the overall performance of precision and recall. The overall accuracy evaluates the overall detection precision of authentic sites and phishing sites. Higher values of P, R, F and O indicate better performance.

We use N_{pp}, N_{ap}, N_{pa}, and N_{aa} to denote the number of phishing sites detected as phishing sites, the number of authentic site detected as phishing sites, the number of phishing sites detected as authentic sites and the number of authentic sites detected as authentic sites respectively.

The detection accuracy of the authentic sites P_1 and phishing sites P_2 are given as follows:

$$P_1 = \frac{N_{aa}}{N_{pa} + N_{aa}},\ P_2 = \frac{N_{pp}}{N_{pp} + N_{ap}} \tag{2}$$

The detection recall of the authentic sites R_1 and phishing sites R_2 are given as follows:

$$R_1 = \frac{N_{aa}}{N_{ap} + N_{aa}},\ R_2 = \frac{N_{pp}}{N_{pp} + N_{pa}} \tag{3}$$

The F-measure of authentic sites F_r and phishing sites F_p are given as follows:

$$F_r = \frac{2 * P_1 * R_1}{P_1 + R_1},\ F_p = \frac{2 * P_2 * R_2}{P_2 + R_2} \tag{4}$$

Meanwhile, the overall detection accuracy O is defined as follows:

$$O = \frac{N_{pp} + N_{aa}}{N_{pp} + N_{ap} + N_{pa} + N_{aa}} \tag{5}$$

4.3 Experiment Design

To evaluate the effectiveness of the proposed method, the Abbasi et al.'s [6] phishing website detection model is chosen as the baseline method. It also consists of many URL and web content features for phishing website detection. Based on these features, the method has a very high accuracy for phishing website detection. However, it doesn't include any domain-specific features, and we can examine whether the incorporation of domain-specific features improves the detection performance. At the same time, we also want to assess the detection performance of the inclusion of genetic algorithm. Thus the experiment consists of two parts. The first experiment is performance comparison between the SMO classification model and Abbasi model, and the second experiment explores the optimization effect of genetic algorithm.

In the first experiment, the collected websites is randomly divided into a training data set and a testing data set. 1023 authentic websites and 991 phishing websites are included in the training data set, while the testing data set consists of 439 authentic websites and 425 phishing websites. The detection precision, recall and F-measure can be calculated for the baseline model and the proposed model without parameter optimization (SMO model).

In the second experiment, we first generated the 10 initial individual genes. Then the individual chromosome was decoded as the value of classification model parameters. Using K cross-validation method, the fitness value of each chromosome is calculated. Iterate the above steps until the best parameters are derived. Based on the derived optimal SMO parameters, the proposed model with parameter optimization (SMO-GA model) and baseline model are trained by a training data set. Then the detection precision, recall and F-measure are calculated for the test data set.

4.4 Data Analysis and Results

At first, we conducted a pair-wise T-test to compare the precision, recall, and F-measures of the SMO model against the baseline model (Table 1). The results indicate that the SMO model significantly outperforms the baseline model across all three performance metrics. These results also illustrate that the proposed context-related feature set results in the higher overall precision in detecting Chinese phishing e-commerce websites than the generic feature sets adopted in the baseline model.

Table 1. The precision (%) comparison of SMO model and Abbasi model

Metrics		SMO model		Abbasi model		MD (M1-M2)
		Mean (M1)	SD	Mean (M2)	SD	
Phishing site	Precision	93.7	0.9	92.0	0. 9	1.7**
	Recall	94.5	0.9	91.0	1.1	3.5**
	F1	94.1	0.4	91.5	0.5	2.6**
Authentic site	Precision	94.6	0.8	91.4	0.9	3.2**
	Recall	93.8	0.1	92.3	0.9	1.5**
	F1	94.2	0.5	91.8	0.5	2.4**
The overall accuracy (O)		94.1	0.4	91.7	0.5	2.4**

SD: Standard Deviation, MD: Mean Difference.
* $p < 0.05$.
** $p < 0.01$.

In order to check whether the genetic algorithm significantly improves the classification accuracy, we conducted a pair-wise T-test to compare the detection performance with and without parameters optimization based on the genetic algorithm. The results shown in Table 2 indicate that the genetic algorithm based parameters optimization significantly improve the performance of authentic websites and phishing websites classification across all three metrics.

Table 2. The precision (%) comparison of SMO model and SMO-GA model

Metrics		SMO model		SMO-GA model		MD (M1-M2)
		Mean (M1)	SD	Mean (M2)	SD	
Phishing site	Precision	93.7	0.9	96.6	0.9	−2.9**
	Recall	94.5	0.9	96.5	0.1	−2.0**
	F1	94.1	0.4	96.5	0.5	−2.4**
Authentic site	Precision	94.6	0.8	96.6	0.1	−2.0**
	Recall	93.8	0.1	96.7	0.9	−2.9**
	F1	94.2	0.5	96.6	0.4	−2.4**
The overall accuracy(O)		94.1	0.4	96.6	0.4	−2.5**

SD: Standard Deviation, MD: Mean Difference.
* $p < 0.05$.
** $p < 0.01$.

5 Conclusion

Developing effective methods for Chinese phishing e-Commerce websites detection has become an urgent task for e-commerce development. However, existing models mainly focus on generic websites classification, which may not be a wonderful solution to detect Chinese phishing e-Commerce websites because they do not consider the specific context-related features in China and face the performance problem. Targeting at detecting Chinese phishing e-commerce websites efficiently, this research incorporates context-related features into the phishing website detection model and adopts the genetic algorithm to determine the optimal classification model parameters. The experiment results show that the context-related features and the parameters optimization method significantly improve the accuracy of Chinese phishing e-commerce websites detection.

There are several limitations of this study. First, we only focus on Chinese phishing e-commerce websites detection. The proposed method needs to be validated in other domains in the future. Second, this study only adopts the genetic algorithm as the parameters optimization method. Considering there are many other artificial intelligence algorithms, it might be interesting to explore the impact of other main artificial intelligence algorithms on parameters optimization.

Acknowledgments. This research is supported by the National Natural Science Foundation of China (Grant No. 71272057, 71572013) and the National Social Science Fund of China (Grant No. 14AZD045).

References

1. iResearch. Annual report of China's E-Commerce (2014). http://report.iresearch.cn/2153.html. Accessed 2014
2. Herzberg, A., Jbara, A.: Security and identification indicators for browsers against spoofing and phishing attacks. ACM Trans. Internet Technol. **8**(4), 36 (2008)
3. APAC. Annual report of Anti-Phishing Alliance of China (2012). http://www.apac.org.cn/gzdt/qwfb/201408/P020140826493067614020.pdf. Accessed 2012
4. Wu, M., Miller, R.C., Garfinkel, S.L.: Do security toolbars actually prevent phishing attacks? In: Proceedings of the 2006 Conference on Human Factors in Computing Systems (CHI 2006), Montréal, Québec, Canada (2006)
5. Ma, J., et al.: Learning to detect malicious URLs. ACM Trans. Intell. Syst. Technol. **2**(3), 24 (2011)
6. Abbasi, A., et al.: Detecting fake websites: the contribution of statistical learning theory. MIS Q. **34**(3), 435–461 (2010)
7. Fu, A.Y., Wenyin, L., Deng, X.T.: Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD). IEEE Trans. Dependable Secure Comput. **3**(4), 301–311 (2006)
8. Zhang, H.J., et al.: Textual and visual content-based anti-phishing: a Bayesian approach. IEEE Trans. Neural Netw. **22**(10), 1532–1546 (2011)
9. Mao, J., et al.: BaitAlarm: detecting phishing sites using similarity in fundamental visual features. In: 2013 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS). IEEE (2013)

10. Huang, H., Qian, L., Wang, Y.: A SVM-based technique to detect phishing URLs. Inf. Technol. J. **11**(7), 921–925 (2012)
11. Gowtham, R., Krishnamurthi, I.: A comprehensive and efficacious architecture for detecting phishing webpages. Comput. Secur. **40**, 23–37 (2014)
12. Bartoli, A., Davanzo, G., Medvet, E.: A framework for large-scale detection of web site defacements. ACM Trans. Internet Technol. **10**(3), 37 (2010)
13. Akiyama, M., Yagi, T., Hariu, T.: Improved blacklisting: inspecting the structural neighborhood of malicious URLs. IT Prof. **15**(4), 50–56 (2013)
14. Xiang, G., et al.: CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. ACM Trans. Inf. Syst. Secur. **14**(2), 21 (2011)
15. Zhang, D., et al.: A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. Inf. Manag. **51**(7), 845–853 (2014)
16. Platt, J.: Sequential minimal optimization: a fast algorithm for training support vector machines. IEEE Trans. Neural Netw. **17**(4), 1039–1049 (1998)
17. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27–38 (2011)
18. Zanni, L., Serafini, T., Zanghirati, G.: Parallel software for training large scale support vector machines on multiprocessor systems. J. Mach. Learn. Res. **7**(3), 1467–1492 (2006)
19. Wang, M., Wang, W.: Approach for kernel selection from SVM ensemble. Comput. Eng. Appl. **45**(27), 31–33 (2009)
20. Kucukkoc, I., Karaoglan, A.D., Yaman, R.: Using response surface design to determine the optimal parameters of genetic algorithm and a case study. Int. J. Prod. Res. **51**(17), 5039–5054 (2013)
21. Rahman, M.A., Islam, M.Z.: A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowl.-Based Syst. **71**, 345–365 (2014)
22. Ilhan, I., Tezel, G.: A genetic algorithm-support vector machine method with parameter optimization for selecting the tag SNPs. J. Biomed. Inform. **46**(2), 328–340 (2013)
23. Khonji, M., Iraqi, Y., Jones, A.: Phishing detection: a literature survey. IEEE Commun. Surv. Tutor. **15**(4), 2091–2121 (2013)