Suspicious URL Detection with Machine Learning

[1]Sanjana Tiwari, [2]Ashok Kumar Behera, [3]Monika Verma

Computer Science Department,
Bhilai Institute Of Technology,Durg

***Abstract*: The web has turned into a key worldwide stage and a basic part of the general public that pastes together day by day correspondence, sharing, exchanging, joint effort, and administration conveyance. Countless associations overall depend on the web for their day by day activities, either totally or just too some degree. These days, the trust in an association vigorously relies upon the nature of its web nearness, which must pass on a feeling of trust and steadfastness to its users over the time. Websites have been created or controlled by attackers for use as attack instruments. A client's machine can be tainted and barged in simply by visiting a pernicious web page, and this sort of attack is known as a drive-by download attack.**

1. Introduction

The web has turned into the mechanism of decision for individuals to search for data, direct business, and appreciate stimulation. In the meantime, the web has likewise turned into the essential stage utilized by scoundrels to attack users. A typical plan to make cash includes the establishment of vindictive programming on countless hosts. Pernicious web content has turned out to be a standout amongst the best instruments for digital culprits to appropriate noxious code. This code is infused into traded off web sites or on the other hand is just hosted on a server under the control of the hoodlums. Whenever a unfortunate casualty visits a noxious web page, the malevolent code is executed and if the unfortunate casualty's program is defenseless, the program is undermined. Subsequently, the injured individual's PC is regularly tainted with malware. As of late, client user has turned into the fundamental target for attacks, as the enemy trust that the end client is the weakest connection in the security chain.A.R Nagaonkar et al 2016 [16]**,** "Finding the malicious URL using search engine mechanism". Author uses different types of method for finding the malicious urls they uses SEO method, link based method, DNS query methods, domain registration methods. Basically author combines lexical and host based features to obtain the accuracy.N. Provos et al 2006,[17] "The Ghost in the Browser," in this paper Author gives the current condition of malware in the internet. The four keypoints outsiders gadgets, promoting, web server security, client contributed substance. All this fetures get combined and used for internet browser services. This paper is only HTML based and JAVA script Based.P. Mavrommatis et al,[18]"All Your IFrame point to us". In this paper only HTML based feature based is used like IFrame.When landing site wants to interact with drive-by-download victim.Client visits the landing site Redirects to get exploit download the malware executable.J Nazario et al[19]"A Virtual Client Honey Pot" In this paper the author examines about a virtual pot. This paper is only HTML based and JAVA script Based.

2. Methodology

In this segment we defines how our proposed system works, Our way to deal with the issue depends on computerized URL classification. Specifically, we investigate the utilization of factual techniques from AI for ordering site notoriety dependent on the connection among URLs and the lexical and host-based highlights that describe them. In contrast to past work, we demonstrate that these techniques can filter through a huge number of highlights got from freely accessible information sources, and can recognize which URL segments and meta-information are significant without requiring overwhelming area ability. Without a doubt, our framework naturally chooses a considerable lot of similar highlights recognized by area specialists as being run of the mill of "malicious" Web destinations. As Figure 2.1shows. To demonstrate this approach, we will built a URL classification system that uses a ICML-2009 datasets.it contains approximately 2.4 million url's and 3.2 million url's features. Using this data,we will extract the features like lexical features(or we can say the printed features of url).lexical features incorporates length of hostnames and url.Host based Features tell "where" is the malicious url destination "their identity" survey by and "How" they are govern. web content based features is the combine feature of host based and lexical based feature. The web content based feature uproot by seize the html page of requested site.it includes HTML count, Hyper link count, Iframe count, Suspicious javascript function count. For training and classification of datasets we use three different classifiers: Linear SVM Classifier, K Nearest Neighbors Classifier, Random Forest Classifier.

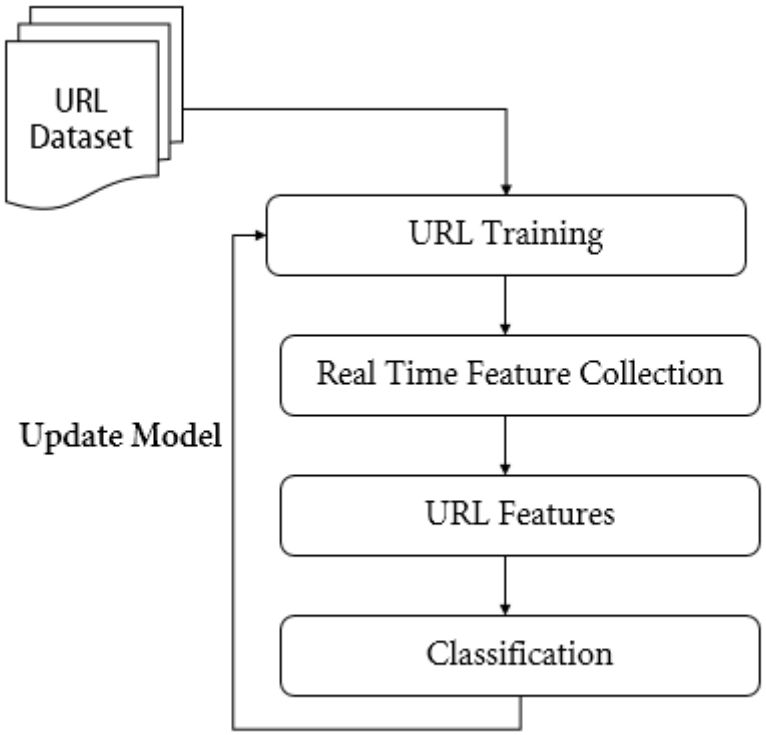


Fig 2.1: Proposed system Architecture

3. RESULT

3.1 Experiment Setup

To evaluate our metrics we have used Python. Data mining tools are used in our evaluation. Project considered various dataset for malicious URL extraction. Table 3.1 shows the specification of our experiment setup.

SNO	Attributes	Value
1	Language	Python
2	Version	3.2 and 2.7
3	Tool	Data Mining Tool
5	Method Used	• SVM Classifier • K Nearest Neighbours Classifier • Random Forest Classifier
6	Dataset	ICML-2009

TABEL 3.1: Experiment Setup Configuration

Run Algorithm to trainee datasets by using svm algorithm. As Figure 3.2shows After that svm algorithm run and provides iterations.

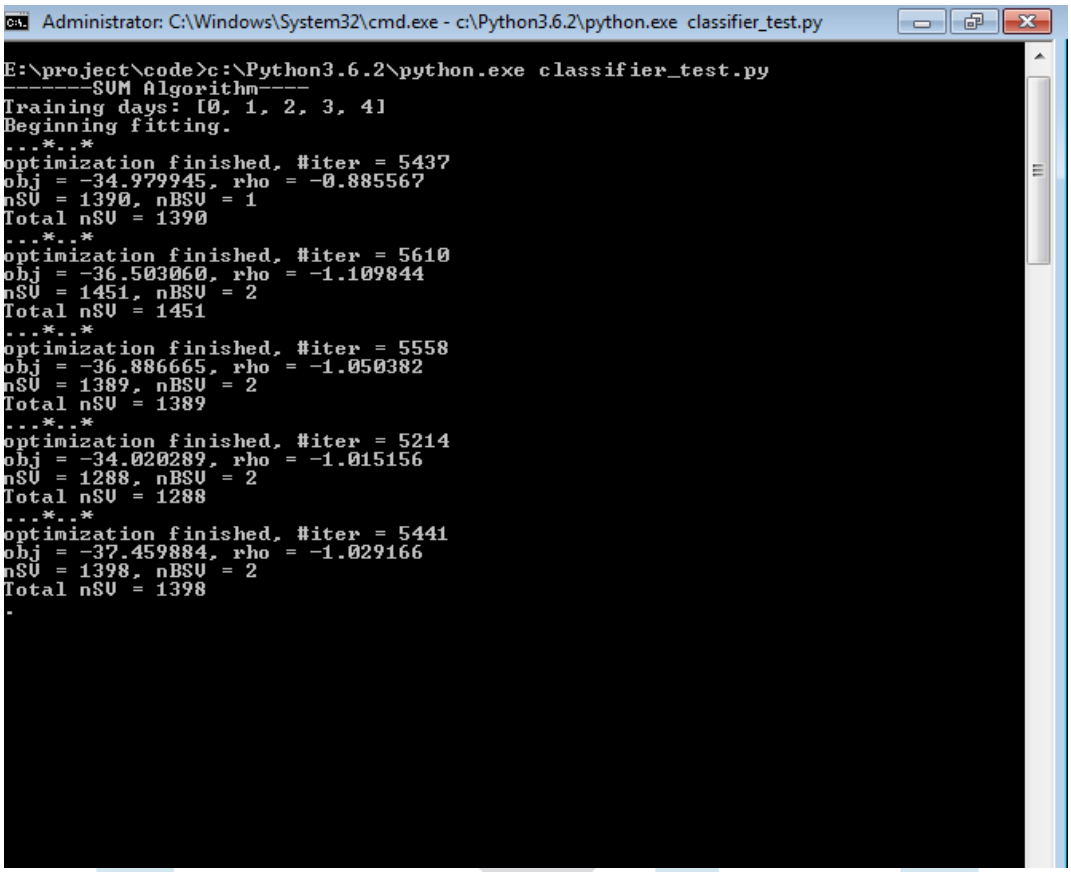


Fig3.2: Training datasets by svm algorithm

Now after completing iterations we achieve accuracy about 97% As Figure 3.5shows by using SVM optimization. By using KNN algorithm we achieve 92% As shown in Figure 3.3, and by using Random forest algorithm we achieve 95% As shown in Figure 3.4.

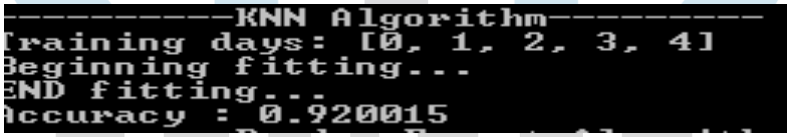


Fig3.3: Training and testing using KNN algorithm

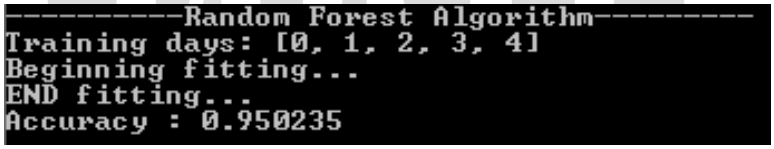


Fig3.4: Training and testing using Random forest algorithm

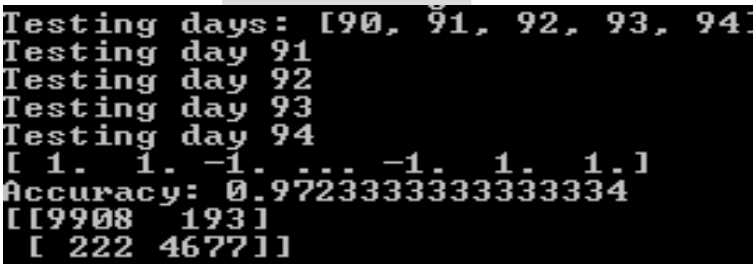
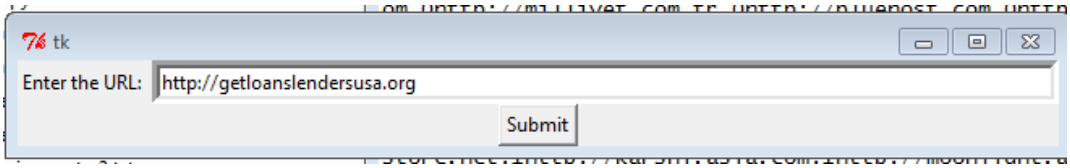
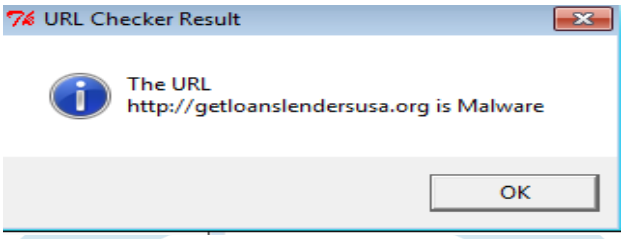


Fig3.5: training and testing using SVM algorithm

Now testing with different urls to find whether it is malicious or not As shown in Figure 3.6.



(a)



(b)

Fig3.6: Testing with URLs-(a) Entering url, (b) Testing url

In this section various experiments are conducted and its results are presented in the form of bar graphs and charts. Python language is used for performing experiments. Project aims to find out the malicious URL's from the given set of URL database. The SVM algorithm is used in combination with the other learning techniques.

4. Conclusion

In the above paper we use the ICMC 2009 data set. We propose a method by utilizing the lexical features, Host based features (IP address, Packets, Token count), Web content based features. By using SVM algorithm we trained and classified datasets by optimizing SVM we get more accurate output then rest. Accuracy produced by Random forest is 95%, K-nearest Neighbor is 92%, and by using SVM we get accuracy of 97%.

5. Future Scope

In this project we use SVM algorithm with ICML 2009 data sets. In future we would like to carry our research to redirect of domain. The Diversion is not traced in this project. Hence malevolent website may contain diversion. In future we will use SVM redirection mechanism.

References

[1] S. B. Rathod and T. M. Pattewar, "A comparative performance evaluation of content based spam and malicious URL detection in E-mail," 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, 2015,
[2] T. Zhang, H. Zhang and F. Gao, "A Malicious Advertising Detection Scheme Based on the Depth of URL Strategy," 2013 Sixth International Symposium on Computational Intelligence and Design, Hangzhou, 2013
[3] L. Fang, W. Bailing, H. Junheng, S. Yushan and W. Yuliang, "A proactive discovery and filtering solution on phishing websites," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015
[4] H. Sha, Q. Liu, Z. Zhou and C. Zheng, "GuidedTracker: Track the victims with access logs to finding malicious web pages," 2014 IEEE Global Communications Conference, Austin, TX, 2014,
[5] M. Akiyama, T. Yagi and M. Itoh, "Searching Structural Neighborhood of Malicious URLs to Improve Blacklisting," 2011 IEEE/IPSJ International Symposium on Applications and the Internet, Munich, Bavaria, 2011, pp. 1-10.
[6] J. Hong, "The state of phishing attacks," *Communications of the ACM*, vol. 55, no. 1, pp. 74–81, 2012.
[7] P. Kolari, T. Finin, and A. Joshi, "Svms for the blogosphere:Blog identification and splog detection," in Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, vol. 4, 2006.
[8] M. Dredze, K. Crammer, and F. Pereira, "Confidenceweighted linear classification," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008,
[9] Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Intelligent rule-based phishing websites classification." IET Information Security 8.3 (2014): 153-160.
[10] Singh, Priyanka, Yogendra PS Maravi, and Sanjeev Sharma. "Phishing websites detection through supervised learning networks." Computing and Communications Technologies (ICCCT), 2015 International Conference on. IEEE, 2015.
[11] Eshete, Birhanu, Adolfo Villafiorita, and Komminist Weldemariam. "Binspect: Holistic analysis and detection of malicious web pages." International Conference on Security and Privacy in Communication Systems. Springer Berlin Heidelberg, 2012.
[12] Zhang, Yue, Jason I. Hong, and Lorrie F. Cranor. "Cantina: a content-based approach to detecting phishing web sites." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.
[13] Miyamoto, Daisuke, Hiroaki Hazeyama, and Youki Kadobayashi. "An evaluation of machine learning-based methods for detection of phishing sites." International Conference on Neural Information Processing. Springer Berlin Heidelberg, 2008.
[14] Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.
[15] Whittaker, Colin, Brian Ryner, and Marria Nazif. "Large-scale automatic classification of phishing pages." (2010).

[16] Jiang, Hansi, Dongsong Zhang, and Zhijun Yan. "A Classification Model for Detection of Chinese Phishing E-Business Websites." PACIS. 2013.
[17] Xiang, Guang, et al. "Cantina+: A feature-rich machine learning framework for detecting phishing web sites." ACM Transactions on Information and System Security (TISSEC) 14.2 (2011): 21.
[18] Thomas, Kurt, et al. "Design and evaluation of a real-time url spam filtering service." 2011 IEEE Symposium on Security and Privacy. IEEE, 2011.
[19] Aburrous, Maher, et al. "Intelligent phishing detection system for e-banking using fuzzy data mining." Expert systems with applications 37.12 (2010): 7913-7921.
[20] Canali, Davide, et al. "Prophiler: a fast filter for the large-scale detection of malicious web pages." Proceedings of the 20th international conference on World wide web. ACM, 2011.