Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210041

Detection of Malicious Web Applications Using Machine Learning Algorithm

Sivagurunathan M[a,1], SivakumarP[b] and Somasundaram. S K[b]
[a]*PG Scholar, Department of Information Technology, PSG College of Technology, Coimbatore, Tamil Nadu, India*
[b]*AssistantProfessor, Department of Information Technology, PSG College of Technology, Coimbatore, Tamil Nadu, India*

Abstract. The emerging development of the internet everyone using web applications for their products or services. The large number of web applications created day by day. Since the demand is very high for web development, developers are creating an application not in a secure manner and hosting without the testing process. Web clients regularly store and oversee basic data that draws in cybercriminals who exploitation the web weaknesses for their benefits. Pernicious website pages are coming to pass undermining issue over the web on account of the reputation and their capacity to impact. Recognizing and examining them is exorbitant due to their characteristics and complexities. The complexities of assaults are expanding step by step in light of the fact that the assailants are utilizing mixed methodologies of different existing assaulting procedures. Using this opportunityattacker used their malicious script in their web application. Attacker, theft user's data, or redirect to malicious websites. In this project, we are proposing detection methods, to prevent the users from approaching the malicious web application. Using a Machine learning algorithm, extract the feature of the web application that is URL features and static features of the network. From the trained model of date set, using the Random forest algorithm detect the malicious web application.

Keywords: Features, Detection, Malicious URL, Machine Learning, Random Forest.

1. Introduction

The development in technology makes the internet grow faster than before. Every day the idea innovates by anyone. At the same time attacker also grow with same technology development. The applications of internet grow reflect everywhere. We are today, it is practically compulsory to have an online presence to run an effective business. The need of internet rapidly growing. Attacker also present in the rapid growth with new attack technique. It is difficult to plan strong frameworks to recognize network safety breaks.

[1]Sivagurunathan M, PG Scholar, Department of IT, PSG College of Technology, Coimbatore, India.
E-Mail id: sivagurunathan.official@gmail.com

A Uniform Resource Locator, [1] known as a URL, is the worldwide location of archives and different assets on the World Wide Web. It is the tool utilized by programs to recover any distributed asset on the web. Fig 1relating a piece of a URL is bolded to exhibit which part is being referred.

Figure 1. Parts of URL

A domain name is an interesting reference that differentiates a site on the World Wide Web. It comes straightforwardly after the protocol and is separated by a colon and two forward slashes. [2] The most popular attack methods are Social engineering and phishing methods and this happen by just visiting the malicious URL of the websites. These attacks happing without finding vulnerability in website, just force the user to click the given link by the attack. [3].Blacklist is basically an information base of URLs that have been declared to be malicious previously. This information base is assembled over the long run, as and when it becomes realized that a URL is malicious. [4] Those techniques are incredibly quick because of a basic inquiry overhead, and henceforth is extremely simple to implement. So, we need use the different algorithm for different attack techniques. [5]

2. Literature Survey

Anand Desai et.al, (2017) proposed, their point is to make an augmentation for Chrome which will go about as middleware between the clients and the harmful sites, and alleviate the danger of clients capitulating to such sites. Further, all hurtful content can't be comprehensively gathered as even that is bound to nonstop turn of events. [6]

Frank Vanhoenshoven et al (2016) used multiple algorithm technique Decision Trees, Random Forest, Naive Bayes, Multi-Layer Perceptron, Support Vector Machines and k-Nearest Neighbours. The mathematical reproductions have demonstrated that most characterization strategies accomplish worthy forecast rates without requiring either progressed highlight choice strategies. [7]

Anton Dan Gabrielet. al (2016) proposes that to present a system that analyzes URLs based on the network traffic and capable of adjusting its detection to new malicious content. Grouped URL is reused as a feature of another dataset that demonstrations as the spine for new recognition models.[8]

Guolin Tanet. al (2018), Unique in relation to the vast majority of past techniques, our work centres on finding vindictive URLs covered inside numerous kind hearted URLs in enormous organization traffic. In this paper, they assess the viability of our methodology on genuine datasets. [9]

Dongjie Liuet. Al (2016), to confirm the adequacy of the strategy, two distinct tests have been directed. To begin with, the proposed strategy was tried dependent on a developed complex dataset. They present correlation results between the proposed strategy and agent AI based recognition calculations. [10-13]

3. Proposed System

Fig 2 represents our proposed methodology, the first step of the implementation is getting the URL and feed in to the dataset. This data set explore and sanitize the URL. After that the URL is feed into real time feature controller that is extract the features from the URL. These features feed up into the classifier. Classifier, classify the necessary details. This detail used in decision trees. Each decision tree has the separate value of each feature. This decision trees combined and make the prediction based upon the score value.

Figure 2. Schematic Representation of Proposed Methodology

3.1 Features Representation

Table 1 represents the lexical features of URL. This is static features of URL, without processing any network side of features. 15 Lexical features listed in table. From this table we can make different type of features trees, that is to predict the URL whether malicious or benign.

Table 2 represents the DNS features of URL, that is based on host-based network details. This can give the more number features to make the decision tree stronger.

Table 1. Lexical Features

Features	Type
Server	Real
Whois register	Real
Resolved IP count	Integer
Name server IP count	Integer
Name server IP count	Integer

Table 2. DNS features

Features	Type
Server	Real
Whois register	Real
Resolved IP count	Integer
Name server IP count	Integer
Name server IP count	Integer

4. Machine Learning Algorithm

Classification is considered as a light weight operation for analyse the URLs whether it is a malicious or non-malicious. Since though the crawling webpage method is considered as a most effective method in reality it come with the time as a cost. In our research we have used most random forest classifier for better classification. [8]

4.1 Machine Learning

Machine learning classifier act as a teacher, which predict probability instance class based on the predefin features. Random forest classifier working based on decision tree nodes. Nodes will be used to find the URL as malicious or not. This decision tree combined together and score system will be calculate. [9,10].

4.2 Flow of Algorithm

The algorithm working in the following way, the RAW file URLs dataset given as the input to the URL predictor. The RAW URLs get into the classifier that separates the URLs as the benign URL and malicious URL. The sanitized URLs are moved into machine learning algorithm that is Random Forest algorithm. This will create the different training data from the sanitized URLs. The training data called as feature extraction that is used for predicting the URL results. The extracted feature given into the training model, this will analysis the features weight and calculate the overall feature value. This value will result the final URLs prediction whether it is malicious or benign.

5. Experimental Results

This is in the form of 80/20 rule. That is initially dataset is divided into training dataset (80%), testing dataset (20%). In order to assess the most efficient mechanism to detect malicious accounts, we inspected various machine learning algorithm. Classifiers are the standard classifiers and widely used in solving problems. So, there are 450175 total numbers of URLs. From that, there is 345737 benign URLs and 104438 malicious URLs. Processed RAW file dataset used into classifier.

The processed dataset is given into classifier, classifier gives the good accuracy detection of malicious URL prediction rate is 98.2 %, the different types of algorithm with accuracy percentage 93.4% and 94.5% for Support Vector Machine and K-nearest Neighbour respectively.

Fig 3 is respective graph of different machine learning algorithm for malicious URL detection. Random forest algorithm gives the 98.2 % of result to the detection method. All types of features processed and the given URL link detected as malicious or benign with high accuracy.

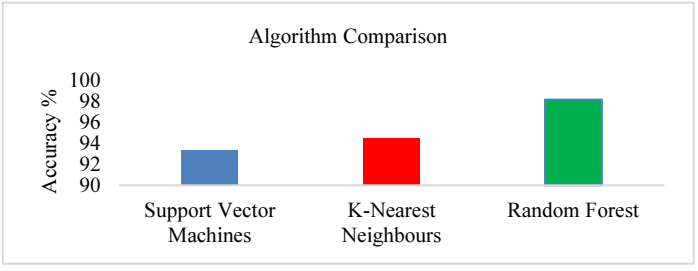

Figure 3. Comparison of Accuracy with proposed & existing Algorithm

6. Conclusion and Future Works

The proposed work analysis provides whether the URL is Malicious or not based on the URL features. The result of the detection method, based upon the decision tree. The algorithm has given a better output after increasing the number of features in the training data. This increases the accuracy of detection. The accuracy which is obtained at 98%, this result is better than the previous works. In future the detection method can be done using various types of new features that depends on exact value.

Reference

[1] Sayambar.A.B, A. M. Dixit, On URL Classification, International Journal of Computer Trends and Technology , 2014.

[2] Merono-Pe˜nuela.A, C. Gu´eret, R. Hoekstra, S. Schlobach et al., Detecting and reporting extensional concept drift in statistical linked data, in 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC. CEUR, 2013.

[3] Xiang et al., A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites, ACM Transactions on Information and System Security ,2011.

[4] Luca Invernizzi, Paola MilaniComparetti, ,EvilSeed: A Guided Approach to Finding malicious Web Pages, Conference Paper, 2012 IEEE Symposium on Security and Privacy, DOI 10.1109/SP.2012.33.

[5] D. R. Patil, J. B. Patil, Survey on Malicious Web Pages Detection Techniques, Science and Technology, 2015 International Journal of u and e- Service.

[6] Anand Desai, JanviJatakia, RohitNaik and Nataasha Raul, 2017, Malicious Web Content Detection Using Machine Leaning, 2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology.

[7] Frank Vanhoenshoven, Gonzalo N´apoles, Rafael Falcon, KoenVanhoof and Mario K¨oppen, 2016, Detecting Malicious Urls Using Machine Learning Techniques, IEEE Symposium Series on Computational Intelligence.

[8] Anton Dan Gabriel, Dragos¸ TeodorGavrilut¸ B˘aetuIoanAlexandru, Popescu Adrian S and tefan, 2016, Detecting Malicious Urls. A Semi-Supervised Machine Learning System Approach, 18thIEEE International Symposium.

[9] Guolin Tan, Peng Zhang, Qingyun Liu, Xinran Liu, Chunge Zhu and Fenghu Dou, 2018, Adaptive Malicious Url Detection: Learning In The Presence Of Concept Drifts, 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering.

[10] Dongjie Liu and Jong-Hyouk Lee, 2016, Cnn Based Malicious Website Detectionby Invalidating Multiple Web Spams, IEEE access

[11] Ramya,T.,Dr.Malathi,S.,ratheeksha,G.R. and Dr.V.D.Ambeth Kumar (2014). Personalized authentication procedure for restricted web service access in mobile phones.Applications of Digital Information and Web Technologies (ICADIWT), 2014, Page(s):69 - 74, Bangalore, India (ISBN:978-1-4799-2258-1)

[12] Ambeth Kumar.V.D (2017).Efficient Routing for Low Rate Wireless Network a Novel Approach. International Journal of Image Mining, Vol. 2, Nos. 3/4, 2017, 2017

[13] Arul E. et.al. (2021) Firmware Injection Detection on IoT Devices Using Deep Random Forest. In: Senjyu T., Mahalle P.N., Perumal T., Joshi A. (eds) Information and Communication Technology for Intelligent Systems. ICssTIS 2020. Smart Innovation, Systems and Technologies, vol 195. Springer, Singapore. https://doi.org/10.1007/978-981-15-7078-0_52