International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Malicious URL Detection using ML

Mrs. Teena Varma[1], Pratik Zinjad[2], Shreeniket Vast[3], Idris Vohra[4], A.Hannan Sunsara[5]

[1]Teena Varma: Professor, Dept. of Computer Engineering, Mumbai University
[2]Pratik Zinjad: Student, Dept. of Computer Engineering, Mumbai University
[3]Shreeniket Vast: Student, Dept. of Computer Engineering, Mumbai University
[4]Idrees Vohra: Student, Dept. of Computer Engineering, Mumbai University
[5]A.Hannan Sunsara: Student, Dept. of Computer Engineering, Mumbai University

---

Abstract: *The internet can be used as a tool for various criminal activities like financial fraud, spam messages or emails for commercial advertisement or can be used for misuse of personal data. All these things are occur due to malicious websites which when opened by user, the data will get to the attacker. Hence precaution before searching any website is necessary. In this paper, we address the detection of malicious URL using machine learning. We will take the sample dataset which contains some malicious URL's and some non-malicious URL's. From the dataset and the use of machine learning algorithm the program can predict that the entered URL or website is malicious or not. It can be useful for security purpose. The aim of this paper is user should know that the searching website is malicious(harmful) or not*

Key Words: Machine learning, Malcious URL, Decision tree, Phisher, Malware.

1. INTRODUCTION

With the growth of Internet usage in the past years, attackers have the advantage of this to hack the client devices using the fake URLs. While by clicking the fake URL's pop ups, we get hacked by the attackers and snippers. Malicious website is common and serious threat to cybersecurity. Malicious URLs are usually websites containing content such as spam, phishing, drive-by exploit email, etc. And normal users become victims of scams such as theft of private information, malware installation and cause losses of billions of dollars every year.

In any phishing attack, the user can trapped into clicking some link to phishing website where user can reveal their sensitive information like username, password, mobile number, email address, etc. Online banking, credit card payments and debit card payments are also become quite popular since past few years. These malicious websites can be used by phishers to get a users card details or bank details. Sometimes user can be trapped to make bank transactions into illegitimate website as well.

In a recent period, attackers and hackers have an advantage using the Internet as their channel to attack the client or users device to gain information using fake URLs. So the users need a method to identify malicious URLs and help them from being victims of scam. So, in this project, we developed the method to identify the malicious and fake URLs with the help of Machine Learning. With Machine Learning algorithms it is possible to teach the machines, to identify the malicious URLs automatically. Malicious URL Detection is an application which will help the users to identify malicious URLs. This application is implemented with the help of a Supervised data using Random Forest Algorithm which uses labeled data to learn how to classify unlabeled data. Malicious URL Detection application helps the user by identifying fake URL. It has features like Checking Protocol, Checking Domain, Checking Path, Checking URL Length, Checking Dash, Checking Dot, Checking Ip Address, Checking Https Token, Checking Web Traffic, Checking URL Date, Checking URL Age, Checking URL DNS, Checking Statistical Report.

The remaining sections of this paper are organized as follows as: Section 2 gives a brief related work. Section 3 describes the methodology with Random forest machine learning algorithm. Section 4 describes the experimental results and discussion. We present our conclusions in Section 5

2. RELATED WORK

In some phishing attacks, a web user can download certain malicious codes, unintentionally. These codes can be javascript codes that are used to attack on a web browser which user using. This attack can results in download of malwares which will harm the various files present in the computer of user. These codes are very difficult to detect. In [1], they have talk about this attack called as "drive by download" attack. In this paper, they have discuss about finding some properties of javascript code that may contains harmful malwares along with it. They work by finding out different features for each of the searched web page. They work on an instrumented web browser to check for the different features for any executed HTML elements or JavaScript codes. By using this method, they were able to get 10 such features of a malicious URL. There

can be different attacking classes which may not contain such features. So, the system will not work in such cases.

In another methodology[2], they make the use of words contained in a URL. Brand names can be placed in the website by the attacker to convince the user for their authenticity. So, in this method, they observe the HTML content of a web page. They perform TD-IDF weighting to give weight to the various words in the HTML content. Then further weighting is performed by considering the website weighting system which helps to get the brand name. This brand name is then get searched with search engine to get domain name of the web page. Then a WHOIS lookup is performed to check that the domain names matched or not. If they match, then the website is considered to be legitimate else not. hence, it mainly makes use of the words that are present in the web page rather than looking the features about the website.

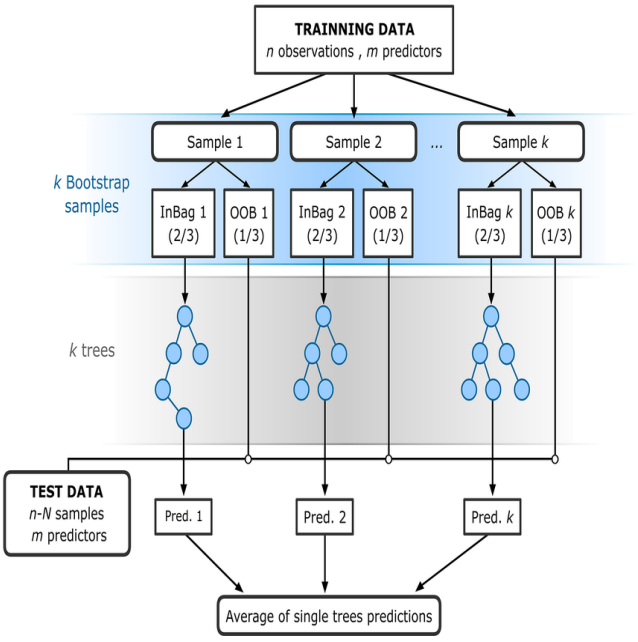
In another method[3],they introduce us to a broader definition of website phishing. This tells how it can be used to trap a user into visiting a new website and reveal their username, password, mobile number and other personal information. It talks about how meta data is a very helpful tool to determine whether a website is phishing or not. Meta data can be made available about a website from the various search engines like Google, Yahoo, Mozilla firefox, etc. This meta data can help in differentiating a phishing website from a non-phishing one. It represent the use of Logistic Regression as a classifier. But it can fail to work when there are a huge number of features and are not necessary that linear in nature.

3. METHODOLOGY

We will take the sample dataset which contains some malicious URL's and some non-malicious URL's. From the dataset and the use of machine learning algorithm the program can predict that the entered URL or website is malicious or not.

The first thing to do after the getting of datasets is data slicing. Here, the datasets are divided into two parts. They are as follows: testing and training dataset. The training datasets are used to train a model. The testing dataset is used once the trained model is ready(for testing purpose). Once the model is trained, we test its accuracy level on the testing datasets. While training the model on the training datasets, we find its accuracy by doing repeated cross-validation on the datasets. This also permit us to do tuning of the parameters in the two datasets to find out which parameter gives the greater accuracy for the datasets. Once, the best suitable parameter for the model is get, the training of the model is done, and then we can go to do testing of the trained model on the testing datasets. We are going to do the classification task on the datasets. We make use of Random forest algorithm.

Random Forest Algorithm: Random Forest algorithm is a supervised machine learning algorithm which can be used to do both regression as well as classification task in data mining. It is an ensemble based method which can be used to perform classification. It makes use of a number of decision trees and after that gives the final result. This algorithm works by creating a large number of decision trees randomly. These trees are created by making use of various samples from the same dataset and also they may use various types of attribute every time to create the trees. Hence, all the trees are made randomly by the use of different sub sets of the same datasets, and also the attributes are taken randomly for the creation of any tree. By doing this, Random Forest Algorithm ensures that it does not over fit the data, as in the case of the classification trees. Once the trees are made, we can do the classification on the basis of result of each tree and then assign it to the class that has been determined by major number of decision trees.



Random forest algorithm flowchart

4. RESULT

Datasets were splited into two parts, training datasets and testing datasets, in the ratio of 70:30. The training datasets were mainly used to train the various models. And then the trained model was implemented on the testing datasets to get the output. While training the model, the repeated cross-validation technique was used to determine the accuracy level of the trained model. The cross-validation

technique was done by creating 10 folds of the training datasets. The method was repeated 3 times. And hence, we get the accuracy output of a trained model. For the tuning of parameters, grid search had been used and repeated cross-validation was did on these values to find the most suitable parameter for a model. This gave the parameter for the model which gives the greatest accuracy on the training datasets. While the testing of this trained model on the testing datasets, to find the accuracy on the testing dataset the confusion matrix was created.

Random Forest was applied on this datasets. Grid search was used to do tuning of its parameter, try. The best value for try was found to be 6 which gave us the highest accuracy on the training datasets. The accuracy on the training dataset by using this value was 93.75%. And when this trained model was applied to testing dataset, an accuracy of 94.11% was get.

Algorithm	Accuracy on training dataset	Accuracy on testing dataset
Random Forest	93.75%	94.11%

5. CONCLUSION

Malicious URL detection is very useful in determining that the certain website is malicious or not and also it should be visited or not. This will help the user a lot in knowing that which of the websites should be avoided. Hence, it will prevent them in revealing their sensitive information to the phisher. It can be very useful for security purpose.

References:

[1] Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection and analysis of drive-by-download attacks and malicious javascript code", Proceedings of the 19th international conference on WWW, pp. 281-290, 2010.

[2] Choon Lin Tan, Kang LengChiew, San Nah Sze, “Phishing Website Detection Using URL-Assisted Brand Name Weighting System”, 2014 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) December 1-4, 2014.

[3] R. B. Basnet, A. H. Sung, "Mining web to detect phishing URLs", Proceedings of the international conference on Machine learning and Applications, vol. 1, pp. 568-573, Dec 2012.

[4] www.google.com

[5]www.geeksforgeeks.com