INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 9, Issue 6 - V9I6-1162)
Available online at: https://www.ijariit.com

Machine learning approach to detect malicious URL using XGBoost algorithm

Dev Kumar
devkr644@gmail.com
Excel Engineering College, Komarapalayam, Tamil Nadu

E. Deepan Kumar
edeepankumar.eec@excelcolleges.com
Excel Engineering College, Komarapalayam, Tamil Nadu

Aarya D. Roy
royaaryaprince@gmail.com
Excel Engineering College, Komarapalayam, Tamil Nadu

Aftab Alam
alam030aftab@gmail.com
Excel Engineering College, Komarapalayam, Tamil Nadu

Harsh Vardhan
harshkrvardhan4@gmail.com
Excel Engineering College, Komarapalayam, Tamil Nadu

ABSTRACT

There are over a billion websites today for the people to visit. People uses the websites to make their works easy but there is a high chance to fall the phishing domain over the internet that inject malware to the client's system or trick them to get their personal details. We will discuss about the machine learning method to classify these URLs in order to prevent people from visiting malicious URLs and improve the security of surfing over the internet. XGBoost algorithm and its performance has been discussed and how it uses the several features of URL to classify and detect the malicious URLs.

Keywords: *URL, malicious URL detection, Machine Learning, Feature extraction, XGBoost algorithm*

1. INTRODUCTION

Nowadays, people use the Internet to simplify their daily activities. Hence, they are likely to access various websites using their URLs (Uniform Resource Locator) in obtaining and sharing information. A report shows that there are more than 1.5 billion websites today, and this number is increasing every second. Unfortunately, all websites are not benign/safe. Hackers often use spam and phishing URL to trick users into clicking malicious URL, the Trojans will be implanted into the victims' computers, or the victims' sensitive information will be leaked. The technology of malicious URL detection can identify malicious URL and prevent users from being attacked by malicious URL.

We have demonstrated how to extract in-depth lexical features, host-based features, and content-based features from URL strings. These features include features extracted from WHOIS, the Wayback machine, statistical features extracted from raw HTML content of a website, and statical features of the URL string itself. The characteristics extracted in this article fall into 3 main categories: Content-Based Characteristics, Host-Based Characteristics, and Lexical Characteristics. This article curates and implements the extraction of these characteristics as URL feature vectors. These X features can be used as feature vectors in malicious URL attribution problems, building predictive models for malicious URL detection or simple fast filters for bad hosts in log streams.

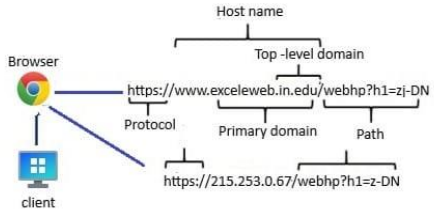
2. RELATED WORKS

The survey in [1] has discussed different methods to detect malicious URLs using machine learning. It showed that the best method among all the methods discussed is machine learning.
In paper [2] five models are used in order to generate new features. The models used were distance metric learning, Nystrom methods DML-NYS, NYS-DML and space transformation models singular value decomposition.
In survey [3] the study of Machine Learning methods- Support Vector Machine, K Nearest Neighbor, Decision Trees, Random Forest, XGBoost, Gradient Boosting, AdaBoost has studied and Neural Network methods- Recurrent Neural Network, Convolutional Neural Network, Generative Adversarial Network, Neural Network Architect has been studied for detecting the malicious URLs.
In [4] the detection of malicious URL by considering the URL features and using Random Forest and Support Vector Machine algorithms. In that it was seen that the Support Vector Machine consider all the features of URL detect the malicious one and URL classification was more accurate when using Random Forest with 100 trees than with 10 trees.
In paper [5] the extraction of 30 URL features were done to identify the phishing domains by using Extreme Machine Learning Algorithm, Support Vector Machine and Artificial Neural Networks classification methods. From these the Extreme Machine Learning Algorithm performed with highest precision.
The extraction of URL features was also done in [6], it mainly focused in the host-based features, lexical features and site popularity features. It used Support Vector Machine and Random Forest algorithm for the detection of malicious domains in which the Random Forest performed better than Support Vector Machine.
In paper [7] the main strategy was to train the machine learning models with large and appropriate data-sets to detect the malicious URLs. It used the Random Forest, Naive Bayes, and Logistic Regression algorithms. The logistic Regression gave the maximum accuracy among the other algorithms.
The main focus in paper [8] is the extraction of URL features and its importance and usefulness for classifying the malicious URLs. The extraction of 18 features were done. It is not very effective classifying short URLs, dark web URLs and embedded URLs.

3. PROPOSED WORK

Understanding the URL

URL is an acronym for Uniform Resource Locator. It is the compact string of numbers, letters and symbols that a computer uses to find a resource on a network and act upon it.
Component of URL:
Protocol - It is a standardized set of rules for formatting and processing data
Hostname - A hostname is part of address typed into a browser to access a website. Traditionally, WWW has been used for web server hostname.
Path - A path is the unique, last part of the URL for a specific function or piece of context.



A malicious URL is a link that is created with the intent of promoting scams and frauds. It contains unsolicited content that can pose a high threat to potential victims.

Features of the URL

Some Lexical Features

Having IP address: To hide domain name attackers use IP address.
Count embedded-domain: Transforms the URL into a unique ID. It can be checked by '//' in the URL.
Suspicious words in URL: Sometimes malicious URLs contain suspicious words such as, login, sign in, account, service, token etc.
Count of https: Commonly the malicious sites do not contain HTTPS protocol.
URL length: Mostly the attackers use long URLs to hide the domain name of URL. The average length of the safe URL is 74.
Google index: In this feature, we check whether the URL is indexed in google search console or not.
Count of digits: If any URL contain digits, it generally indicates as suspicious URL because safe URLs doesn't contain digits. So, the number of digits in URLs is also an important feature.

Some Host-Based Features

ISP: Connection speed to host
TTL: Last Updated date
Registration Country: Country of registration
Hosting Country: Country of hosting

Open ports: Ports open on host

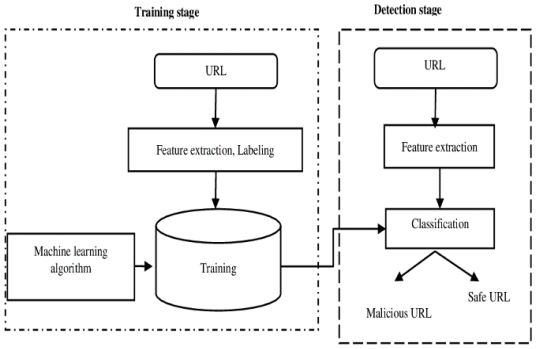
Number of open ports: Total number of open ports

Some Content-Based Features

Number of HTML tags: Total number of HTML tags on page.
Number of Hidden tags: Total number of tags with class or id as 'hidden or attributes of visibility or display as none'.
Number of iframe: Total number of iframe tags on page
Number of objects: Total number of objects tags on page
Number of embeds: Total number of embed tags on page
Number of internal hyperlinks: Total number of internal hyperlinks on page
Number of external hyperlinks: Total number of external hyperlinks on page

Model

We have used the XGBoost algorithm to detect the malicious URLs. The model will be trained and tested by using URL data-set having 6,51,191 URLs. The dataset consists of 4,28,103 benign URLs, 32,520 malware URLs, 96,457 defacement URLs and 94,111 phishing URLs.
XGBoost builds a tree model by using gradient boosting framework to provide parallel tree boosting, that can be further divided into Sub-models. Its Parameters specify real number arguments for selected objectives which make the boosting process more conservative and prevents overlifting.



Feature extraction and Labeling

The lexical, host-based and content-based features will be extracted from the URLs that will help identify the structure of the URLs and its properties for the accurate classification. These extracted features will be the input for training the model. The type of URLs is labeled as 0,1,2,3 for benign, defacement, malware and phishing URLs respectively. For the new URL input the model will extract it's features first and then classify the type of URL. The use/count of these lexical features in the URL is almost same for the same type of URLs.

Training and Testing

The data-set needs to be split into two parts with the ratio of 80:20 for the training and testing of XGBoost model.
The model will be trained by passing the 80 percent of the data-set as the training data and the rest 20 percent of the data-set will be used to test the model accuracy.

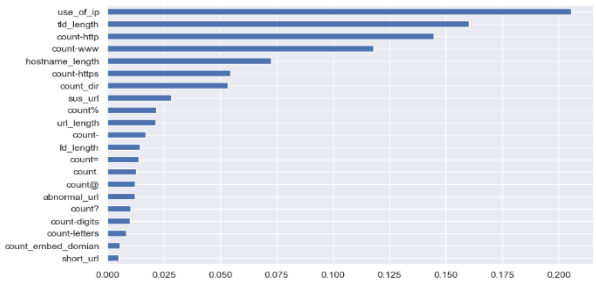


Figure 1. Feature importance graph

The importance of URL features is shown in figure 1. It shows which feature is most and least important or mostly considered for detecting the malicious URL using the XGBoost model.

Detect type of new URLs

The input string of URL will be taken from the user and then the features are extracted from the URL to classify its type using XGBoost model.

4. RESULTS AND DISCUSSION

The confusion matrix in figure 2 shows the performance of the model with predicted values and actual values. It is clearly mentioned how many predictions were correct and wrong.

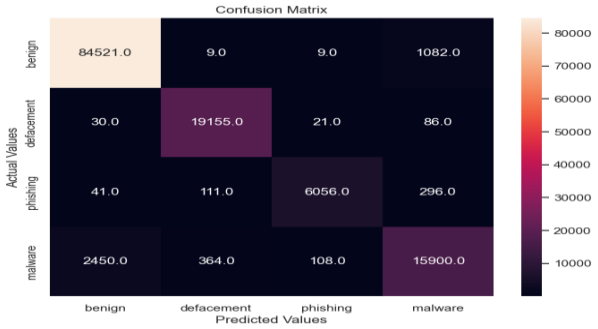


Figure 2. confusion matrix

Accuracy: it is the percentage of correct predictions.

Acc = (TP + TN /(TP+TN+FP+FN))*100

Precision: precision = (TP/(TP+FP))*100
Recall: recall = (TP/(TP+FN))*100
Where, TP is true positive: the prediction was positive and it is actually positive.
TN is true negative: the prediction was negative and it is actually negative.
FP is false positive: the prediction was positive and it is actually negative.
FN is false negative: the prediction was negative and it is actually positive.
F1-score: the model having high F1-score is considered to be good classification model.
F1 = (2*(precision*recall))/(precision + recall)

	precision	recall	f1-score	support
benign	0.97	0.99	0.98	85621
defacement	0.98	0.99	0.98	19292
phishing	0.98	0.93	0.95	6504
malware	0.92	0.84	0.88	18822
accuracy			0.96	130239
macro avg	0.96	0.94	0.95	130239
weighted avg	0.96	0.96	0.96	130239

accuracy: 0.965

Figure 3. XGBoost performance

The precision, recall and f1-score for all four types of URL is mentioned in figure 3.
The XGBoost model has 96 percent of accuracy for the 80:20 split of data-set for training and testing data. We can improve the accuracy of the model by different split ratio of the data-set.

5. CONCLUSION

In this paper, we have discussed about the lexical, host-based and content-based features extraction of the URL and how it is used to classify and identify the malicious URLs. We have demonstrated the machine learning approach to build malicious URL detection model using XGBoost algorithm and seen the importance of URL features in detecting the phishing domain and the performance of XGBoost algorithm. It has obtained 96 percent of accuracy in detection of malicious URLs.

6. REFERENCES

[1] Lekshmi A R, Seena Thomas (2019) "Detecting Malicious URLs Using Machine Learning Techniques: A Comparative Literature Review", International Research Journal of Engineering and Technology (IRJET).
[2] Tie Li, Gang Kou, Yi Peng (2020) "Improving Malicious URLs Detection via Feature Engineering: Linear and nonlinear Space Transformation Methods", Information Systems (Elsevier).
[3] Eint Sandi Aung, Hayato Yamana, (2020) "Malicious URL Detection: A Survey", Department ofomputer Science and Communication Engineering, Graduate School of Fundamental Science and Engineering.

[4] Cho Do Xuan,Hoa Dinh Nguyen, Tisenko Victor Nikolaevich,(2020) "Malicious URL Detection based on Machine Learning", International Journal of Advanced Computer Science and Applications.
[5] Yasin Sonmez, Turker Tuncer, Huseyin Gokal, Engin Avci (2018) "Phishing Web Sites Features Classification Based on Extreme Learning Machine", 6th International Symposium on Digital Forensic and Security (ISDFS)
[6] Ripon Patgiri, Hemanth Katari, Ronit Kumar and Dheeraj Sharma, (2020) "Empirical Study on Malicious URL Detection Using Machine Learning", International Conference, ICDICT.
[7] Vanitha N and Vinodhini V, (2019) "Malicious URL Detection using Logistic Regression Technique", International Journal of Engineering and Management Research.
[8] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma, (2019) "Detection of Malicious URLs using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE).

