URL Filtering by Using Machine Learning

Malik Najmus Saqib

Department of Cybersecurity, College of Computer Science and Engineering
University of Jeddah
Saudi Arabia

Summary
The growth of technology nowadays has made many things easy for humans. These things are from everyday small task to more complex tasks. Such growth also comes with the illegal activities that are perform by using technology. These illegal activities can simple as displaying annoying message to big frauds. The easiest way for the attacker to perform such activities is to convenience user to click on the malicious link. It has been a great concern since a decay to classify URLs as malicious or benign. The blacklist has been used initially for that purpose and is it being used nowadays. It is efficient but has a drawback to update blacklist automatically. So, this method is replace by classification of URLs based on machine learning algorithms. In this paper we have use four machine learning classification algorithms to classify URLs as malicious or benign. These algorithms are support vector machine, random forest, n-nearest neighbor, and decision tree. The dataset that is used in this research has 36694 instances. A comparison of precision accuracy and recall values are shown for dataset with and without preprocessing.

Keywords:
Input here the part of 4-5 keywords.

1. Introduction

Each year technology has been growing to new extend and dimensions. Such growth has affected the way the businesses promote themselves. It starts from simple marketing strategy to a complex web-based services. Such services host multiple servers. These websites are assessable by using Universal Resource Locator (URL). Similarly, the tremendous growth in technology has evolved the method of from simple to sophisticated attacks. Few examples of such attacks are fake websites, thief of digital money by manipulating user to reveal their credentials, identity theft or installing a backdoor on a victim system. Many techniques have been discovered to launch such attacks for last decade. These techniques are Man in the Middle (MITM), SQL injections, website phishing, social engineering, distributed Denial of Service and so on. The security technologies are also evolving to cater such malicious techniques. However, the exponential growth of risks, threat and security attack is becoming a serious concern.

The attacking techniques on the information system are mainly based on malicious URL. Such URL when goes unnoticed by the end user causes a serious damage to the individual or organization. A URL consists of two major parts: a protocol identifier and a resource name. A protocol identifier is used to indicate which protocol are used. A resource represents the domain name of the resource location. There are many characteristics feature of URL. There are some characteristics that are used to identify the website as malicious. Most of those features are textual.

A URL is said to be malicious if it possesses security threats of any type. Many URL nowadays are malicious [1], that lead to the website that contains exploitable content like worms, backdoor, phishing, spam and etc. User that are not aware of malicious URL threat become the victim very easily.

It is very important that the user should be given awareness about the possibility of becoming a victim of security threat via malicious URL. Also, a comprehensive counter measures should be applied to detect malicious URL. Initially, blacklist is used to detect malicious URL. This method has been deployed by many antivirus companies for years. Blacklist is basically a collection of URLs that has be found malicious. This collection is updated as soon as new malicious URL is found. However, the attacker uses sophisticated methods like obfuscation to bypass blacklist by making a URL that look like a legitimate. There mainly four techniques of Obfuscation [2]; using IP address to obfuscate host, using wrong name for host, using another domain name for the host and using big host name for the host. With these techniques, attacker hide the malicious URLs [2].

Many techniques have been used to detect malicious URL. These approaches can be categorized as Listing based approach and Machine learning based approach. Listing based approach is the classical approach that has been used for many years. In this approach a list of malicious URLs is maintained. any request to a URL would first scan the list, if the new URL is present in the list, then it is malicious otherwise not. The biggest limitation of list-based approach is that it is almost impossible to update the list with the newly coming malicious URLs [3]. Thus, it is easy for the attacker to bypass he list by generating new malicious URLs.

Manuscript received August 5, 2022
Manuscript revised August 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.8.34

Nevertheless, the listing-based techniques is used by many antiviruses' software nowadays.

The paper is organized as follow. Related literature review is described in section 2. Section 3 discusses the preliminaries of our work, that includes the structure of URL, and the threats possess by malicious URL. Section 4 detailed the research methodology of this research work. The experiments and results are compared in section 5 and section 6 conclude the paper.

2. Related Work

In this section of the paper, we have discussed state of the art literature review that uses machine learning techniques to detect malicious URLs.

In [4], the researcher proposed a malicious URL filtration technique that uses the machine learning approach. This research performed the comparison of various machine learning techniques that are support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and K-Nearest Neighbor. With these machine learning algorithms, they also used Artificial Neural Network (ANN). Their dataset consists of various features, but they used only those features to train the classifiers that are most important.

The efficiency of filtering malicious URLs via machine learning approach is discussed in [5]. A bio inspired algorithm is used with the features optimization approaches to detect malicious URLs. They have used a static analysis technique for machine learning. The optimization of URLs features are performed by using bio inspired algorithm called Particle Swarm Optimization (PSO). It is concluded in this research that using Support Vector Machine (SVM) and Naïve Bayes shows good results with high accuracy of detection. The detection accuracy rate that is achieved is 99%.

In [6], the researcher has built up the URL from the benign and malicious websites. A numeric representation is used to represent the features of URL. least absolute shrinkage and selection operator (LASSO) and Multi-objective Pareto Genetic algorithm (MOGA) are the two methods that are used to pick up the most relevant features. A classifier based on Support Vector machine is trained by using the URLs with the selected features. The performance of this classifier is evaluated by using a ten-fold validation method. A comparison study is performed among the two feature selection methods used. The achievable performance is more that 95% accuracy of classification and F-scorer of the selected most relevant features.

A strengthening technique of the current approaches is proposed in [7]. This strengthening technique uses machine learning approach. The set of hosts based, and lexical feature is fourteen. But the focus is on the employment of typical methods for detection of malicious URLs. The main machine learning approaches used by the proposed mechanism are Random Forest and Decision Tree. The proposed mechanism is trained and tested on set of benign and malicious URL dataset. The proposed mechanism provided the accuracy of more than 97%.

Researchers in [8] has proposed and developed a technique to detect phishing attack with only 9 most requirable features. They concluded that these nine features are enough to detect phishing attacks. They have used ISCXURL-2016 dataset for training and testing their technique. In this dataset there are 11964 benign and malicious URLs records. They have used various machine learning classifiers to see which classifier gave maximum accuracy. It is concluded that 99.57% accuracy is achieved by using the Random Forest algorithm.

A merger of linear and non-linear space methods is proposed in [9] to improve classification models that filter the malicious URLs. A two-layer distance metric approach for learning is used for linear transformation. Nyström method is used for nonlinear transformation. A dataset consists of 331,622 URLs is used to test the proposed method. It has 62 various features. The performance and efficiency of various classification methods has been improved significantly. The classifiers are k-Nearest Neighbor, Support Vector Machine, and neural networks. An increased from 68% to 86% has been observed to identify the malicious URL in k-nearest Neighbor. An increase from 58% to 81% is seen in the rate of detection by using Linear Support Vector Machine.

3. Preliminaries

In this section we have discussed various parts of URL and the possible threat from the malicious URL. Such threat if not avoided or prevented become a successful attack.

2.1 Parts of URL

There are different parts of URL. Each part describes a specific thing. The detail description of URL is described in RFC 3986 [12].

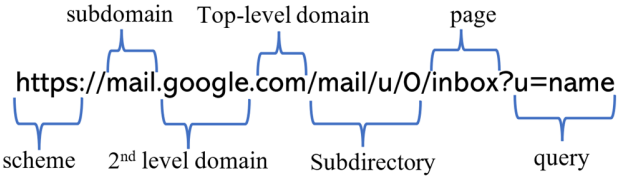

Fig 1: Parts of URL

Fig. 1 shows various parts of URL, which are described below:

- Scheme: it specifies the protocol used by the web server to access the resource from the website. Hypertext Transfer Protocol Secure (HTTPS) is the most common and preferable scheme used nowadays.
- Subdomain: it represents that which specific page from the whole website is being accessed.
- 2nd level domain: It is the actual name of the website. Through this, visitors know which company/organization website they are visiting.
- Top-level domain: it represents the type of registration of the visited website on the Internet. E.g. ".edu" represents the academic organization and ".com" represents the commercial organization.
- Subdirectory: it represents a folder on the web server. It helps to understand that in which specific section of the website you are currently visiting.
- Page: It usually comes after the sub folder path. It tells the visitor that which page he/she is accessing. The server will return default page in case it is missing.
- Query: It represents the query that will be executed on the server side. It is separated by a question mark symbol.

2.2. Possible threat from malicious URL

Attacker prepare the malicious content that he wants to be executed for user. Attacker can keep those malicious content in the defaced legitimate website or hosting his own website. The purpose of the attacker is to make the user to click on the malicious link so that the threat able content is executed. The attacker must trick the user to click on the malicious link in a defaced website or to visit the malicious URL of the attacker website.

If the website is hosted by the attacker for malicious content only then the attacker is free to use any URL for malicious content. In such case he can even use the default page of the website for malicious content. And that would be hard to detect by the security mechanism incorporated.

4. Research Methodology

This section of the paper discusses the various part of our experiments. This section describes the dataset used in experiment, feature extraction and classification algorithms used for detection malicious URLs.

3.1 Dataset

The dataset we use is ISCXURL2016 [11]. Data set contain 36694 instances. The format of dataset file is Comma Separated Values (CSV) where each record is in one row and each value in every record is separated by comma. Table 1 shows the detail of the URL in the dataset with various categories.

Table 1: URL dataset categories

Type	***A4 Paper***
Phishing	7575
malware	6711
spam	6698
Defacement	7930
Benign	7780

Table 2 shows the various fields in the selected dataset and their description, The first step is to scale of the dataset because each instance in the dataset has different ranges. If the scaling is not done, then it will be problematic for the classification model. We used weka tool for a data scaling approach for each of independent variables so that all variables are on similar scale. We apply unsupervised filter normalize to scale the minimum and maximum value of data in a similar range.

Table 2: URL dataset fields

Dataset filed	***Explanation***
domain_token_count	No. of token in a URL
executable	URL is pointing to executable file or not
NumberofDotsinURL	Count the number of dots (.) in URL
Arguments_LongestWordLength	Count the character in the largest word in URL
NumberRate_Domain	Occurrence of domain name in dataset
NumberRate_FileName	
NumberRate_AfterPath	
Entropy_Domain	Domain entropy
class	Classification of URL

3.2 Selection of feature of URL

A crucial step in any machine learning system is the identification of the features. It is very important to select

the features of high significance for machine learning. A sharply tuned features are very helpful in categorizing the website as malicious or benign. We have selected the feature so that we get best results of our selected classification algorithm, which are discuss in the next section.

3.3 Classification

There are few classification approaches of machine learning used in this research work. A comparison study of those approaches for malicious URL detection for the selected dataset is performed in the next section.

Random Forest is a famous classification algorithm. This algorithm is a supervised learning algorithm. It works by converting the dataset into several sub decision trees and then merges those sub trees to obtained better prediction and accuracy. The decision tree's hyper parameters and a seizing classifier are used by Random Forest algorithm. The subset of YRL's features is used to split the node of the tree. Arbitrary threshold values make the decision sub tree random.

Another approach that we used for classification is Decision Tree. Decision Tress is machine learning algorithm that make use of decision supports methods. It is a supervised and non-parametric learning algorithm. It allows to use sub tree in the decision process. The trees are converted into edges based on the important element that is known as conditional tree node. In this algorithm, the decision is the leaf node of the tree, that is last node in the branch, and it cannot be split.

A famous supervised learning model is Support Vector Machine (SVM). This learning algorithm works by observing data and finding appropriate patterns. Suck patterns are used for analysis and classification. This algorithm is uses on the multidimensional space for making hyperplanes. These hyperplanes categorize cases into various classes. This learning algorithm is capable of handling multiple variables of various categories. Support Vector Machines are capable to give accurate results if the number of samples in the dataset is less than the number of dimensions of dataset. Therefore, it is also called high-dimensional spaces. This algorithm is adoptable in many environments and uses memory efficiently. This algorithm might give poor results if the number of samples is much less than the number of features.

Another famous supervised learning and non-parametric machine learning algorithm is k-Nearest Neighbor (kNN). This algorithm makes classes of separate data point for prediction and classification by using the concept of proximity. Basically, it uses a database. In this database the samples are divided into various groups. These groups help in predicting the category of new data sample. There is not assumption of data made in this model. This machine learning algorithm can be used for classification of data and for regression process. But mostly it is famous for classification of data.

5. Experiments and Results

We use the tool Weka 3.8.6 for the experiment. The dataset that is used in the experiment is mentioned in sub-section 3.1. The first process is the cleanness of the dataset. This process is also discussed in sub-section 3.1.

The dataset consists of 9 features, which are mentioned in Table 2. These 9 features are used for the classification. We have calculated the performance of Random Forest, Decision Tree, Support Vector machine and k-nearest neighbor algorithms. Table II shows the results without preprocessing of the dataset while Table shows the results after preprocessing of dataset.

Table 2: Performance analysis without preprocessing

Classifier	Accuracy	Precision	Recall
Random Forest	0.914	0.936	0.93
Decision Tree	0.855	0.927	0.929
SVM	0.897	0.874	0.846
K-NN	0.923	0.939	0.899

It can be seen that the performance of Random Forest classification algorithm is almost same as the performance of k-nearest neighbor algorithm.

Table 3: Performance analysis with preprocessing

Classifier	Accuracy	Precision	Recall
Random Forest	0.949	0.974	0.93
Decision Tree	0.915	0.961	0.92
SVM	0.927	0.935	0.956
K-NN	0.933	0.969	0.949

6. Conclusion

In this research work show machine learning algorithm for the classification of malicious URL and benign URL. The dataset chosen is consists of 36694 URLs. In this dataset 78.7% URLs are malicious and 21.3% URLs are benign. We have compared the four classification algorithms which are Random Forest, Decision Tree , Support Vector machine and k-nearest neighbor. The performance comparisons of these four classification algorithms are shown in table 2 and table 3.

Acknowledgments

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-02-042-DR). The authors, therefore, acknowledge with thanks the University of Jeddah technical and financial support.

References

[1] Liang, B., Huang, J., Liu, F., Wang, D., Dong, D., and Liang, Z., *Malicious Web Pages Detection Based on Abnormal Visibility Recognition*. In EBISS. IEEE (2009)

[2] Garera, S., Provos, N., Chew, M., Rubin, A., *A framework for detection and measurement of phishing attacks*. In Proceedings of the 2007 ACM workshop on Recurring malcode. (2007)

[3] Sheng, S., et. al. *An empirical analysis of phishing blacklists*. In Proceedings of Sixth Conference on Email and Anti-Spam (CEAS) (2009).

[4] A. Bhagwat, K. Lodhi, S. Dalvi and U. Kulkarni, "*An Implemention of a Mechanism for Malicious URLs Detection*," 2019 6th International Conference on Computing for Sustainable Global Development, 2019, pp. 1008-1013

[5] Lee, O. V., Heryanto, A., Razak, M., "*A malicious URLs detection system using optimization and machine learning classifiers"*, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 17, No. 3, March 2020, pp. 1210~1214

[6] G. Chakraborty and T. T. Lin, "*A URL address aware classification of malicious websites for online security during web-surfing*," 2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2017, pp. 1-6, doi: 10.1109/ANTS.2017.8384155.

[7] Tung, S., Wong, K., Kuzminykh, I., Bakhshi, T., and Ghita, B., "Using a Machine Learning Model for Malicious URL Type Detection", In Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 21st International Conference, Russia, August 26–27, 2021.

[8] Gupta, B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., Chang, X., "*A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment*", Computer Communications, Volume 175, Pages 47-57, 2021ondal, D., Singh, B., Hu, H., Biswas, S., Alom, Z., Azim, M., "*SeizeMaliciousURL: A novel learning approach to detect malicious URLs*", Journal of Information Security and Applications, Volume 62, 2021

[9] Li, T., Kou, G., Peng, Y., "*Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods*", Information Systems, Volume 91, 2020.

[10] Canadian Institute for Cybersecurity, University of New Brunswick, URL dataset (ISCX-URL2016)

[11] RFC 3986 Uniform Resource Identifier (URI): Generic Syntax, Network Working Group, 2005 Online: https://datatracker.ietf.org/doc/html/rfc3986