DETECTING MALICIOUS URLS USING MACHINE LEARNING TECHNIQUES: A COMPARATIVE LITERATURE REVIEW

Lekshmi A R[1], Seena Thomas[2]

[1]M.Tech Student in Computer Science and Engineering at LBS Institute of Technology for Women, Trivandrum, Kerala.

[2]Associaste professor in Computer Science and Engineering department at LBS Institute of Technology for Women, Trivandrum, Kerala.

---***---

Abstract - *Today the most important concern in the field of cyber security is finding the serious problems that make loss in secure information. It is mainly due to malicious URLs. Malicious URLs are generated daily. This URLs are having a short life span. Various techniques are used by researchers for detecting such threats in a timely manner. Blacklist method is famous among them. Researchers uses this blacklist method for easily identifying the harmful URLs. They are very simple and easy method. Due to their simplicity they are used as a traditional method for detecting such URLs. But this method suffers from many problems. The lack of ability in detecting newly generated malicious URLs is one of the main drawbacks of Blacklist method. Heuristic approach is also used for identifying some common attacks. It is an advanced technique of Blacklist method. But this method cannot be used for all type of attacks. So this method is used very shortly. For a good experience, the researchers introduce machine learning techniques. Machine Learning techniques go through several phases and detect the malicious URLs in an accurate manner. This method also gives the details about the false positive rate. This review paper studies the different phases such as feature extraction phase and feature representation phase of machine learning techniques for detecting malicious URLs. Different machine learning algorithms used for such detection is also discuss in this paper. And also gives a better understanding about the advantage of using machine learning over other techniques for detecting malicious URLs and problems it suffers.*

Key Words*: *Blacklist, Cyber Security, Malicious URL.

1. INTRODUCTION

The growth and promotion of businesses spanning across many applications including online-banking, e-commerce, and social networking due to the advent of new communication technologies. The use of the World Wide Web has increasing day by day. By using the Internet, most of the time malicious software, shortly named malware, or attacks are propagated. Delivering malicious content on the web has become a usual technique for bad actors due to increased internet access by more than half of the world population. The explicit hacking attempts, drive-by exploits, social engineering, phishing, watering hole, man-in-the middle, SQL injections, loss/theft of devices, denial of service, distributed denial of service, and many others are variety of techniques used to implement website attacks. The limitations of traditional security management technologies are becoming more and more serious given this exponential growth of new security threats, rapid changes of new IT technologies, and significant shortage of security professionals. By spreading compromised URLs most of these attacking techniques are realized.

Malicious URLs are compromised URLs that are used for cyber-attacks. To avoid information loss the URL identification is the best solution. So the identification of malicious URL is always a hot area of information security. Drive-by Download, Phishing and Social Engineering, and Spam are most popular types of attacks using malicious URLs. Downloading of malware by visiting a URL is referred as Drive-by-download. By exploiting vulnerabilities in plugins or inserting malicious code through JavaScript, Drive-by-download attack is usually carried out. For affecting the genuine web pages, the phishing and social engineering attacks ploy users into disclosing their sensitive information. Spam is used as voluntary messages for the purpose of advertising or phishing. Every year this types of attacks leads several problems. So the main concern exists today is detecting such malicious URLs in a timely manner.

In this paper we mainly discuss about the different methods used for detecting malicious URLs. This paper mainly focus on the advantage of machine learning techniques in the field of detecting malicious URLs over other techniques. The classification of feature representation stage of machine learning techniques is explained in detail. The methodical classification different feature extraction stage of machine learning techniques is well explained. The paper provide a good explanation about different machine learning techniques used for detection. Moreover this paper gives an additional information about the difficulties caused by user in using machine learning techniques for the malicious URL detection.

2. METHODS OF DETECTING MALICIOUS URLS

The main methods for detecting malicious URLs are blacklist method, heuristic method and machine learning method. These are explained as follows.

2.1 Blacklist Approach

This is the most common method currently used for malicious URL detection. This is one of the classical method for detecting malicious URLs. According to Jian Zhang et al. [1] the blacklist method is having a database consists of a set of URLs that are malicious in past. This URLs that are malicious in past. This method is very fast and easy to implement. Whenever a user visits into a new URL, then a search for database is performed in blacklist. If the new URL is already a member in the blacklist then a warning will be generated to show that the URL is malicious otherwise the URL is a benign one. This technique can have very low false positive rates. The attackers making complications on Blacklist Approaches.

S. Ganera et al. [2] identifies mainly four types of complications in this approach. The first one is that using an IP it makes confusion on host. Second reason for complication is confusing the host with other domains. Third one is making confusion on host by using large host names. Final one is vitiate. A new complications in Blacklist approach is found by the recent increase in the significant extent of URL in a considerable amount i.e., the URL is made substantially shorter and direct into a required page. Y. Alshboul et al. [3] discovers this new complication on blacklist approach that the battering of malicious URL beyond a short URL. Fast-flux and algorithmic generation of new URLs are some other techniques used by attackers to avoid the blacklist approach.

S. Sinha et al. [4] proposes a reputation-based blacklist approach. Here the compromised hosts as well as malicious contents in URLs, network, and host can be identified. Then using this information the access of web, emails and other activities in malicious networks or host can be blocked. Many organizations like intrusion detection, spam detection uses this type of approach. A SpamAssassin and DSpam are two spam detector used in this approach for detecting malicious URL in user's mail. Manual training is needed for DSpam for detection. A number of spam detectors are used by SpamAssassin and scores of each detectors can be assigned. According to his view a Searching for databases in blacklist are done by IP address reversing, blacklist zone can be appended and then a DNS lookup is done. But there occurs many problems in this approach. The most important drawback of this approach is maintaining comprehensive list of malicious URLs is very difficult.

The blacklist approach is useless of predicting newly generated new URLs. According to S. Sheng et al. [5] view it is impossible to detect new ultimatum in daily generated URLs. The other drawbacks of Blacklist method are which halts the signature based tools from detecting the attacks by complicating the code. More attacks are bring out by the attackers and which amend the attack signature.

2.2 Heuristic or Rule-based Approach

The blacklist method can have the lack of ability to detect newly generated malicious URLs. So a supplement method is used by C. Seifert et al. [6] that is the Heuristic approach. This method is mainly used for identifying the phishing sites by extracting phishing site features. This method creates a blacklist of signatures whenever a new URL is arrived. Then it is analyzed with the available list of signatures. If there is a match exists, then the URL is referred as malicious. In this approach, a signature is assigned to each of the recognized common attacks based on its behavior. Two methods are used here to detect malicious URLs using heuristic approach. They are Signature based method and Behavior based method. The signature based detection method is described in P. Gutmann [7]. It is like fingerprint property i.e., it's an eccentric property. From malicious sites, different patterns are extracted. This type of method can have small error rate. But this method can have a major disadvantage that it requires additional amount of time, money and work force for bring out distinctive signatures.

The behavior based detection method is described in [8]. Based on the behavior property it is concluded that whether the URL is malicious or benign. Here the URLs with some same behavior are collected for detection. Because of using the system resources and services in a similar manner, these type of technique can be used in detecting URLs generated in a deviant form. According to G. Jacob et al. [9], this method have a data collector, an interpreter and a matcher. The data collector gather together both static and dynamic information for execution. Interpreter is used for translating data collection module into intermediate representation. Matcher examine the above representation with behavior signature. The main goal of this method is to detect the exotic and distinct malicious variants. But it uses large amount of time for scanning, and it doesn't contains details about false positive ratios. Data mining techniques as well as machine learning techniques are used Heuristic approach to acquire knowledge about the executable files behavior. In M. Schultz et al. [11] describes the heuristic approach uses Nave Bayes and Multi Nave Bayes for classifying the URLs as malicious or benign one. Nave Bayes is a classification algorithm for binary and multi class classification problems. While working on a dataset with millions of records with some attributes this type of classification algorithm is used. Nave Bayes method is from Bayes theorem. This classifier assumes that all features are unrelated to each other. i.e., the presence or absence of a feature does not influence the presence or absence of any other feature. But this algorithm has a disadvantage that it cannot learns the relationship between features since it consider all the feature to be unrelated.

The Nguyen et al. [12] explains heuristic approach in his paper. According to his view, the heuristic based detection technique analyses and extracts phishing site features and detects phishing sites using that information. Then extract features of URL in user-requested page and applies those features to determine whether a requested site is a phishing site. It helps to reduce damage caused due to phishing attacks.

According to Nida Khan1 et al. [13], there are two module for heuristic approach for detecting malicious URLs. In this approach, the URL and DNS matching module is the first module. This module contains a white-list like blacklist method. The increase in running time and decrease in false negative rate can be managed by the use of white-list. It contains two parameters: domain name tie in with IP address. When a user enters into a website, then the system finding a match with domain name of current website and white-list. If the website present in white-list then it can be accessed by user, then for checking DNS poisoning attack the system equates IP address of corresponding domain. At the beginning the white-list start with zero. Because there is no domain in the list. When a user access new web page or enters a new URL, then the white-list start increasing. The most important point is that while the website or URL accessed by the user at the first time, the URL domain will not present in the white-list. The second module is starts from identification module. This module detect the phishing web page or URL by extracting hyperlinks of web page and applying the phishing detection algorithm such as ID3 or C4.5 and finally a decision is taken such as whether the URL of the web page is malicious or benign. If it is malicious a warning will be generated. But the disadvantage of heuristic approach is that it uses large amount of system resources and the whenever visit in a website an immediate attack is launched. Due to this the generated attack may go undetected.

2.3 Machine Learning Approach

The drawbacks of blacklist method and heuristic approach is discussed in the previous two sections. To overcome these problems, most of the researchers started applying machine learning techniques in detecting malicious URLs from benign one. From [14] the machine learning can have high demand on Artificial Intelligence (AI). These technique provide a system to learn by itself and improve from experience without having any specific program. The main goal of machine learning is to provide ability to the computer to learn automatically without any interference with humans. From [15] the important machine learning techniques are supervised earning, unsupervised learning and semi supervised learning. The supervised learning have a teacher for giving guidance. This method take some data set and act as a teacher. It guides the model or give training to the machine. After proper guidance it starts take decisions on the arrived new data. In unsupervised learning, they can learn without the help of a teacher. I.e., from observations and some structure of data. The working of this method is different from supervised learning. This method create clusters by finding patterns as well as relationships from a given set of data. Semi supervised learning [14] is act between supervised and unsupervised learning. For training it takes both labeled and unlabeled data. This method of machine learning gives the better performance and high accuracy. The machine learning techniques does the following [15].

1. To create a model, the various machine learning algorithm is trained by using a set of training data.
2. Once a new input data is entered into the machine learning algorithm it takes some prediction on the basis of the trained model.
3. The prediction taken in step 2 can be evaluated for checking accuracy.
4. If the estimated accuracy is once tolerable, then the machine learning algorithm is deployed. Otherwise using an enhanced set of training data the machine learning algorithm is again and again trained.

D. Sahoo et al. [10] states that for training the data, machine learning techniques uses a set of URLs and for classifying a URL as malicious or benign this method learns a prediction model. This is mainly based on the statistical property of URLs. The main advantage of this type method is that the ability to specify new URLs.

3. MALICIOUS URL DETECTION: ALGORITHMS AND PHASES

3.1. Batch learning algorithm and online learning algorithm: The main algorithms used in machine learning technique to detect malicious URLs are batch learning and online learning. The batch learning in machine learning has an assumption that before the training task the training data is available. SVM (Support Vector Machine) algorithm and Nave Bayes algorithm are the two main families of batch learning algorithm. SVM [17] is a supervised learning method. It can make full use of structural risk minimization principle with a maximum margin learning approach. The main aim of SVM is to find a hyper plane for classifying the data points in N dimensional space. So the main concern keep in mind while finding a hyper plane is the hyper plane can have maximum margin. That means the malicious class and benign class can have maximum distance between the data points. The hyper plane depends on the number of input features. If the number of input feature is 2, then the hyper plane is just a line. If the number of input feature is 3, then the hyper plane becomes 2-dimensional. But it is very difficult to imagine a hyper plane with more than 3 input features. Due to this for considering the whole features of a URL, SVM produces less accuracy. Nave Bayes [16] is an important classification algorithm. It takes the tested URLs as input and it gives the output that the testing domain names with their attack type. It can do the following steps.

1. Using the training set, it estimate the sub-features for training purpose and using some Gaussian distribution, a classifier is created.
2. Estimate the mean and variance of sub features found in step 1.
3. For classification, it takes the calculated features for testing sample.
4. Calculate the anatomy for benign, spam, malware classes.
5. Examine the values of anatomy.
6. For testing domain, highest value obtained in step 6 can be assigned.

This method gives more accuracy than SVM for all features of a URL. Because in SVM considering the whole features of URL makes some complication. So it may affect in whole performance. But there is no such issues in nave Bayes method. They can improve performance of the entire technique and gives more accuracy. Online learning algorithms [18] are used for handling data with high volume data high velocity. It's a family of scalable learning algorithms. This algorithm is mainly trying to find a solution for classification problems. During training, this type of algorithm make a label prediction at each step. Then they receives an actual label. Perceptron algorithm, Logistic regression with SGD (Stochastic Gradient Descent), and ConfidenceWeighted (CW) algorithm are some family of algorithms included in online learning. The perceptron algorithms are classical algorithms. It update the weight vector whenever a malicious URL is detected. It is a linear classifier algorithm. The rule of update in perceptron algorithm is very simple. This algorithm cannot have any chance for misclassification since it update rate are fixed. So this method gives better results for detecting malicious URLs. That is the false positive ratio is low here. The Logistic Regression with SGD is an online method for providing an approximation of gradient descent used in batch learning algorithms. This method express the sum over individual examples. So this algorithm is gives the details of all false positives and false negatives in detail. It increases the accuracy. But its performance is least when compared to perceptron algorithm. For each feature, the CW algorithm maintains a different confidence measure. Here more confidence weights are more submissive than less confident weight. In contrast with other algorithms it describes per-feature confidence with a Gaussian distribution. So the classification happens here in an efficient manner. The accuracy of performance is high. The advantages of online learning algorithm when compared to batch learning is that its computation speed is high because in this algorithm makes one pass over the entire data. Since the batch learning algorithm possess multi-pass on data. Due to the one pass behavior of online learning, this type of algorithms are easy to implement. Sometimes the online learning makes hard to learn. Because for several cases most of the algorithms cannot act correctly or may be it misbehaves. Both the batch learning algorithm and online learning algorithm go through two phases of machine learning to detect malicious URLs. They are feature extraction and feature representation. The machine learning technique first extracts all features of a URL, then represent those features and then implement the above algorithms.

3.2. Feature extraction phase: The Feature Extraction Phase is to extract the different features of a URL. From Anjali et al. [16] the feature extraction module mainly consists of six features. They are lexical features, link popularity features, webpage content features, network features, DNS features and DNS fluxiness features. The lexical features is mainly based on URL string or URL name properties. The length of URL, length of each component of URL such as host name, top-level domain and primary domain, the number of special characters in URL are the most commonly used statistical properties of URL string. These statistical properties are included in traditional lexical features. In this feature a dictionary is constructed for each types of words. So each of this word become a feature. A BOW (Bag-Of-Words) model is used in this feature for showing the value of feature as 0 or 1.i.e., while the word in the dictionary is present in URL, then the feature value shows 1 otherwise 0. According to [10] some advanced lexical features are also used. It include directory related feature like length of directory and number of subdirectory, file name features like length of file name and the number of delimiters and finally the argument features such as length of arguments and number of variables.

From [16], the link popularity means the total amount of incoming links from possible websites. Link popularity features helps to identify malicious URL and benign one. Because the benign URL can have large amount of link popularity when compared to the malicious one. Malicious URLs only possess a few amount of link quality. In this method, the URLs domain link popularity and address link popularity is employed.

Host based features [10] are very important in detecting malicious URL detection. This feature includes properties of IP address, WHOIS information, properties of domain name, location information and connection speed.

DNS fluxiness features [16] are include in host based features. This DNS fluxiness features are used for proxy network identification and host changing. This property is helpful for identifying URLs fluxiness nature. Webpage content feature is included in content based features. This type of features are heavy weighted. Here the main concern is to extend lot of information within a safely manner. The content based features gives the information about the threads in earlier even if the detection of malicious URLs using URLs features fails. The webpage content features have the ability to count the number of HTML tags, iframe, lines and hyperlinks in webpage.

DNS features and Network features are referred by the authors in [16]. DNS features related to the name of an address. It includes the count of resolved IP. Network features gives the information about network such as length of downloaded packet content, number of bytes in downloaded file, speed of download and domain lookup time.

3.3. Feature representation phase: The Feature Representation phase [10] includes features collection stage and feature pre-processing stage. The feature collection stage collects the most important information about a URL. The presence of URL in blacklist, URL string information, host information, HTML content and JavaScript content of the website, link popularity are included in this phase. The feature pre-processing phase contains the URLs textual description. i.e., the URLs unstructured information. This information can fed into the various machine learning algorithms by converting to a numerical vector.

4. CONCLUSION

Today the detection of malicious URL is very important. The cybercrimes are increased day by day due to the unknown malicious URLs usage in cyber security field. Various type of methods are used in detecting this type of attacks. Among this, machine learning technique is a favorable solution for such attacks. In this survey different types of methods used for detecting malicious URLs are discussed and also identified the drawbacks of each method and advantages of machine learning over all other methods. Even though machine learning techniques are better and suitable solution for detecting malicious URLs.

REFERENCES

[1] Jian Zhang, Phillip A. Porras, and Johannes Ullrich. Highly predictive blacklisting. In Proceedings of the 17th USENIX Security Symposium, July 28-August 1, 2008, San Jose, CA, USA, pages 107122, 2008.
[2] S. Garera, N. Provos, M. Chew, and A. D. Rubin, A framework for detection and measurement of phishing attacks, in Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007, pp. 18.
[3] Y. Alshboul, R. Nepali, and Y. Wang, Detecting malicious short URLs on twitter, 2015.
[4] S. Sinha, M. Bailey, and F. Jahanian, Shades of grey: On the effectiveness of reputation based blacklists, in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 5764.
[5] S. Sheng, B. Wardman, G. Warner, L.F.Cranor, J. Hong, and C. Zhang, An empirical analysis of phishing blacklists, in Proceedings of Sixth Conference on Email and AntiSpam (CEAS), 2009.
[6] C. Seifert, I. Welch, and P. Komisarczuk, Identification of malicious web pages with static heuristics, in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 9196.
[7] P. Gutmann. The Commercial Malware Industry., 2007.
[8] KALPA, Introduction to Malware, 2011.
[9] G. Jacob, H. Debar, and E. Filiol, Behavioral detection of malware: from a survey towards an established taxonomy, Journal in Computer Virology, pp. 251266, 2008.
[10] Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi, Malicious URL Detection using Machine Learning: A Survey, 2017 International Conference on. IEEE.
[11] M. Schultz, E. Eskin, E. Zadok, and S. Stolfo, Data mining methods for detection of new malicious executables. In IEEE Symposium on Security and Privacy, pages 38-49. IEEE COMPUTER SOCIETY, 2001.
[12] Nguyen, Luong Anh Tuan,”A novel approach for phishing detection using URL-based heuristic.” Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.
[13] Nida Khan1, Manliv Kaur1, Riddhi Panchal1, Prashant Kumar Rai1, Nilesh Rathod2 Heuristic Based Approach for Fraud Detection using Machine Learning International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 2, February 2018.
[14] Marco Varone, Daniel Mayer, Andrea Melegari https://www.expertsystem.com/machinelearning-definition/
[15] Atul Published on Jan 09, 2019 https://www.edureka.co/blog/what-ismachine-learning/
[16] Anjali B. Sayamber, Arati M. Dixit, Malicious URL Detection and Identification International Journal of Computer Applications, August 2014.
[17] M. A .Hearst, S. T. Dumais, E. Osman, J. Platt B. Scholkopf, Support Vector Machines, Intelligent systems and their applications, IEEE.vol.13, no.4.pp.18-28,1998.
[18] Justin M A, Berkeley Lawrence K. Saul, Stefan Savage and Geoffrey M. Voelker, Learning to Detect Malicious URLs ACM transactions on intelligent systems and technology, vol. 2, no. 3, article 30, publication date: April.