IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 10, October 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010012 |

Malware Detection Using Machine Learning

Manjushree Kabade, K Arjun

PG Student, Dept. of C.S.E., Bheema Institute of Technology & Science, Adoni, India

Asst. Professor, Dept. of C.S.E., M.Tech(PhD), Bheema Institute of Technology & Science, Adoni, India

ABSTRACT: Cyber-attacks against sensitive data have become one of the serious threats. New malware or malicious programs are released everyday by cyber all over the world due to the rising applications of computer and information technology criminals through the Internet in an attempt to steal or destroy important data. The aim of these attacks is to steal the information used by individuals and organizations to conduct transactions. Malware Link's websites contain various hints among their contents and web browser-based information. . In order to cope with new variants of malicious software, machine learning techniques can be used for accurate classification and detection. The proposed work is an efficient scheme for malware detection and protecting sensitive data from malicious threats using machine learning technique. The purpose of this study is to detect the malicious website by applying the Random Forest Algorithm on URL dataset.

I. INTRODUCTION

Malware is a term used for any malicious software which enters the system without authorization of user of the system. By emerging the words "Malicious" and "software" the term is created as Malware. Most of the malware enters the system while downloading files over internet. Malware has ability to infect the other executable code, data/system files, boot partitions of drives and create excessive traffic on network leading to denial of service. When user executes the infected file, it becomes resident in memory and infect any other file executed afterwards. Malicious software or malware breaches the secrecy and integrity of data and causes unauthorized leakage of information.

In recent years, cyber-attacks are increasing alarmingly because of the emerging applications of computer and Internet. In today's age, it has become almost necessary to have an online presence to successfully run venture. As a result, the importance of World Wide Web has been continuously increasing. Since most of the users go online to access the services provided by government and financial institutions, there has been a significant increase in phishing attacks for the past few years. The hackers have started to earn money and they are doing this as a successful business. The reason for creating these websites is to get private data from users like account numbers, login id, passwords of debit and credit card, etc.

To improve the generality of malicious URL detectors, Machine learning algorithms have been one of the most powerful techniques introduced. The proposed work is a systematic approach for malware detection for protecting sensitive data from malicious threats using machine learning techniques.

Objective

A framework for malware detection aiming to detect the malicious URL using Random Forest Algorithm is used. Other automate classification algorithms could also be used in this framework. After having been successfully tested on medium-size datasets of malware and clean files, the ideas behind this framework were submitted to a scaling-up process that enable us to work with very large datasets of malware and clean files.We propose a versatile framework in which one can employ different machine learning algorithms to successfully distinguish between malwarefiles and clean files, while aiming to minimize the number of false positives.

Scope of the study

MALWARE is defined as software designed to damage a computer system without the owner's informed consent. Malware is actually a generic definition for all kind of computer threats. Malware detection through standard, signature based methods is getting more and more difficult. since all current malware applications tend to have multiple

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010012 |

polymorphic layers to avoid detection or to use side mechanisms to automatically update themselves to a newer version at short periods of time in order to avoid detection by any antivirus software.

II. METHODOLOGY

1. First the user will randomly select the URL instance from the dataset.
2. The selected URL instance undergoes the malware detection technique.
3. Random Forest algorithm is applied on the instances for the detection of malware.
4. In the feature extraction, 30 features are extracted from the URL instance given.
5. For each feature output will be either 1, 0, -1 and these details are stored in csv file.
6. Random forest classifier uses this as training data and predict whether the URL instance is malware or legitimate.

Malware and its classification

Malware is defined as "any code added, changed or removed from a software system in order to intentionally cause harm or subvert the intended function of the system". The fact that malware can cause loss of information, money as well as life represents a big threat to technology advancements. The classification of malware depends on execution characteristics of the program. Malware is also classified depending on its payload, how it exploits or makes the system vulnerable and how it propagates.

Malware analysis technique

1. **Analysis using dynamic methods:** In Dynamic methods for malware analysis we execute the file and collect the information about its properties like what actually is it's intend. he advantage these dynamic methods is that we can make sure that what actually will happen when this type of malware's in an actual system
2. **Analysis using static methods:** A static method of analysis of malware is like using the fix patterns on source code and detect is behavioral properties Analysis using static methods have an advantage that with good accuracy they can detect its purpose and functionality

URL

URL is the abbreviation Uniform Resource Locator, which is the global address of the documents and other resources on the World Wide Web. A URL has two main components such as Protocol identifier (indicates what protocol to use) and Resource name (specifies the IP address or the domain name where the resource is located).

URL Analysis

Malicious URLs can be analyzed based on the lexical features and host based features of the URL and the structures. The lexical feature analyses the format of the URL an URL consists of two parts the hostname and the path.

Types of attacks using malicious URLs include:

1. **Drive-by Download**: Drive-by download refers to the (unintentional) download of malware upon just visiting a URL. Such attacks are usually carried out by exploiting vulnerabilities in plugins or inserting malicious code through JavaScript.
2. **Phishing and Social Engineering**: Phishing and Social Engineering attacks trick the users into revealing private or sensitive information by pretending to be genuine webpages.
3. **Spam**: Spam is the usage of unsolicited messages for the purpose of advertising or phishing. These attacks occur in large numbers and have caused billions of dollars' worth of damage.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010012 |

Python for Malware Detection

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- Python is Interpreted - Python is processed at runtime by the interpreter. We do not need to compile our program before executing it. This is similar to PERL and PHP.
- Python is Interactive - we can actually sit at a Python prompt and interact with the interpreter directly to write our programs
- Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- Python is a Beginner's Language - Python is a great language for the beginner-level programmers and supports the development of a wide.

Python's features include

- **Easy-to-learn** − Python has few keywords, simple structure, and a clearly defined syntax. This allows the user to pick up the language quickly.
- **Easy-to-read** − Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** − Python's source code is fairly easy-to-maintain.
- **A broad standard library** − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable** − we can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases** − Python provides interfaces to all major commercial databases.
- **GUI Programming** − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** − Python provides a better structure and support for large programs than shell scripting.

Anaconda prompt

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010012 |

Advantages of using anaconda prompt

- Anaconda includes Python plus about 600 additional Python packages
- Anaconda installs without administrator privileges
- Anaconda makes package management and virtual environments easier
- Machine learning algorithms for malware detection

SYSTEM ANALYSIS AND DESIGN

The system analysis is necessary to examine the flow of documents, information, and material to design a system that best meets the cost, performance, and scheduling objectives.

Existing System

Disadvantages:

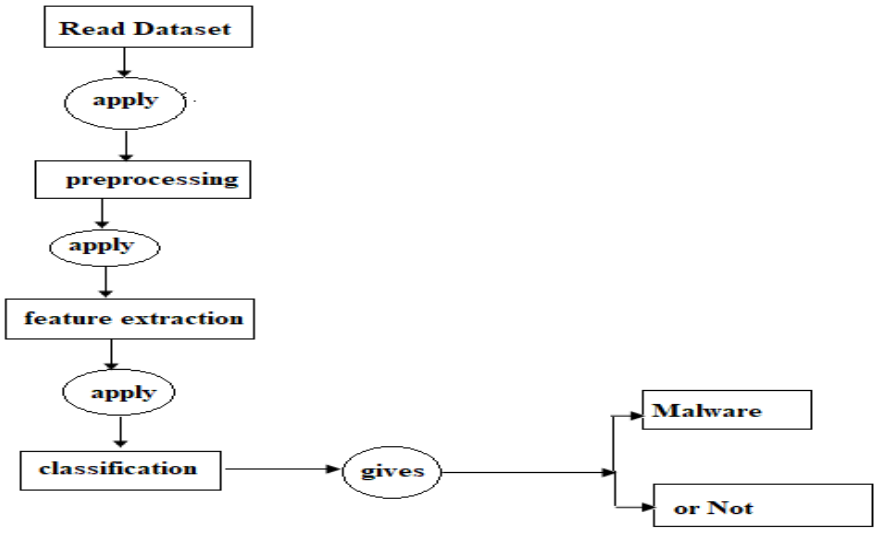
- Unable to detect new malware
- Easy to bypass
- Require update database frequently
- As the social Malware attacks underscore the dangers of the public it takes all the personal information's and need to adequate countermeasures.
- In existing methods they fail to find the Malware websites, but they tried it to a mark upto 50% still they can't succeed.

PROPOSED SYSTEM:

Advantages:

- This study is considered to be an applicable design in automated systems with high performing classification against the Malware activity of websites.
- Furthermore, in literature comparisons, this study is observed to be high-performing by having a high performance.

Design Methodology



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010012 |

Machine learning algorithms for malware detection

Decision tree

Tables Decision tree is like a flow chart diagram consisting of root node, leaf node and branches. Root node is the starting point of the tree and leaf nodes are the condition based results; branches are arrows connecting to node shows flow from question to answer. Internal nodes corresponding to attribute and leaf nodes corresponds to class labels Decision trees helps obtaining the result in faster and in an efficient way. The two types of decision tree created are, Classification trees: In this target variable is categorical and tree is used to identify the class Regression tree: target variable is continuous and tree is used to predict its value.

Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. **Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.**We use random forest because, It takes less training time as compared to other algorithms and predicts output with high accuracy, even for the large dataset it runs efficiently.

Working of Random Forest Algorithm

The working of the algorithm consists of following steps:

Step 1: First, start with the selection of random samples from a given dataset.

Step 2: Next, the algorithm will construct a decision tree for every sample then it will get the prediction result from every decision tree.

Step 3: Voting will be performed for every predicted result.

Step 4: Select the most voted prediction as the final prediction result.

The Advantages of using Random Forest are,Random Forest is capable of performing both Classification and Regression tasks. It is capable of handling large datasets with high dimensionality and it enhances the accuracy of the model and prevents the overfitting issue. Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

IMPLEMENTATION

To run our application the first step is to use the anaconda prompt to navigate local host in our browser. After the completion of local host server connection, we will get a web page which consists of an option as Enter a url and a submit button. Once the user enters the URL and press the submit button, the entered url undergoes the malware detection to find whether it is malicious or legitimate.

III. RESULTS AND DISCUSSIONS

Anaconda prompt application with the python programming language is used for the implementation process. Py server.py command is used to navigate local host in our browser. Once the connection is established, we will get a web page consisting of a Enter a URL option and a submit button. When the user enters the URL in the given space and press submit button. The entered URL undergoes the implementation steps such as data preprocessing, feature selection and classification. An output of array 0, 1, -1 is obtained. Based on this outcomes the prediction is done. If the outcome is -1 then we consider the URL has malware and if the outcome is 1 then we consider the URL as legitimate.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 10, October 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1010012 |

IV. CONCLUSION AND FUTURE SCOPE

Our main goal was to develop a machine learning framework that generally detects as many samples of malware as possible. The random forest algorithm used was able to predict the behavior of URL whether it is malware or legitimate, by giving the highest accuracy of 96.777 for the URL instances.

Future Scope:

We will explore further correlations between Malware sites and hosting and DNS registration companies. We will also look at additional features that can be leveraged, such as Content Security Policies, certificate authorities, and TLS Finally, we will look at the underlying HTML structure for features, specifically counts of tags, placement of tags, use of and counts of specific JavaScript functions, inline and included CSS etc.

REFERENCES

[1] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Compute. Appl., vol. 25, no. 2, pp. 443–458, 2014.

[2] R. M. Mohammad, F. Thabtah, and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," Internet Technol. …, pp. 492–497, 2012.

[3] W. Hadi, F. Aburub, and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," Appl. Soft Comput. J., vol. 48, pp. 729–734, 2016.

[4].G. Canbek and "A Review on Information, Information Security and Security Processes," Politek. Derg., vol. 9, no. 3, pp. 165–174, 2006.

[5].L. McCluskey, F. Thabtah, and R. M. Mohammad, "Intelligent rulebased phishing websites classification," IET Inf. Secur., vol. 8, no. 3, pp. 153–160, 2014.

[6].Flow-based malware detection using convolutional neural network M. Yeo ; Y. Koo ; Y. Yoon ; T. Hwang ; J. Ryu ; J. Song ; C. Park 2018 International Conference on Information Networking (ICOIN) Year: 2018 | Conference Paper | Publisher: IEEE

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details