MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

DATA MINING FOR MALICIOUS CODE DETECTION SYSTEM

[1]Mr. Mahesh N Gunjal, [2]Ms. Ketaki R Takawale,
[3]Mr. Akshay B Raut, [4]Ms. Sonam P Jadhav,
[5]Prof. Vinod Wadne

Department of Computer Engineering and Technology,
JSPMs Imperial College of Engineering & Research,
Wagholi,Pune, India.

[1]maheshgunjal31@gmail.com, [2]ketakitakawale@gmail.com,
[3]technoakshay@gmail.com, 4sonamj.jadhav@gmail.com
[5]vinods1111@gmail.com

Abstract: *Data mining is the process of posing queries and extracts the patterns into the unknown from the previously located large quantities of data using pattern matching or other reasoning techniques. These patterns can be seen as details of the input data, and may be used for to analyzing in machine learning and predictive analytic of data. Data mining has many applications in security including for national security as well as for cyber security. The threats to national security include hijacking destroying critical infrastructures such as power grids and telecommunication systems also the E-commerce. Cyber security is involved with protecting the computer and network systems against corruption due to Trojan, malware, spyware, worms and viruses. Data mining is also being applied to provide solutions such as intrusion detection and auditing. Data mining is also becomes other applications include data mining for malicious code detection such as worm detection and managing firewall policies. The various types of threats to national security and describe data mining techniques for handling such threats. Threats include non real-time threats and real-time threats [5].Data mining is also being applied for credit card fraud detection and biometrics related applications. Another challenge is to mine multimedia data including surveillance video. Finally, we need to maintain the privacy of individuals. The root kit records were categorized as Inline and other based on the attribute values. In this paper, we proposed three algorithms [10] named as RIPPER [10], Navies Bayes approach, and Multi-Naive Bayes using data mining techniques and the comparison of these algorithms. The various types of threats and then discuss the applications of data mining for malicious code detection, cyber security and national security.*

Keywords: *Malicious Code Detection, Data Mining, Computer Security, Prediction, Machine learning.*

1. INTRODUCTION

Data mining techniques are being used to sort out the individual or groups that are capable of doing some these types of the terrorist tasks. Malware refers to a broad class of malicious software that threatens computer and information systems and networks. The goal of intrusion detection is to discover intrusions into a computer or network, by observing various computer network activities or attributes, given the interceptive growth of the Internet and the increased availability of tools for attacking networks, intrusion detection becomes a critical component of network administration. Root kits are known as software that is used to hide the presence and activity of malware (such as viruses, worms and Trojans) and allow an attacker to take control of a system. Such software may modify, destroy or stolen data may obtain unauthorized access to confidential data and exploit vulnerabilities in applications. Whereas the most appropriate examples of such type of devices so far are smart phones, notebooks and tablets, which are more powerful than early personal computers and can be used easily. Hence, a reliable and fast automation of the analysis and classification is a crucial point to be able to cope with this threat. The advantage of the proposed system over the existing system is to develop a system that provides misuse detection of intruders based on signature analysis. The proposed Intrusion Detection system will be running on the server as compared to Existing system shown in Fig.1.

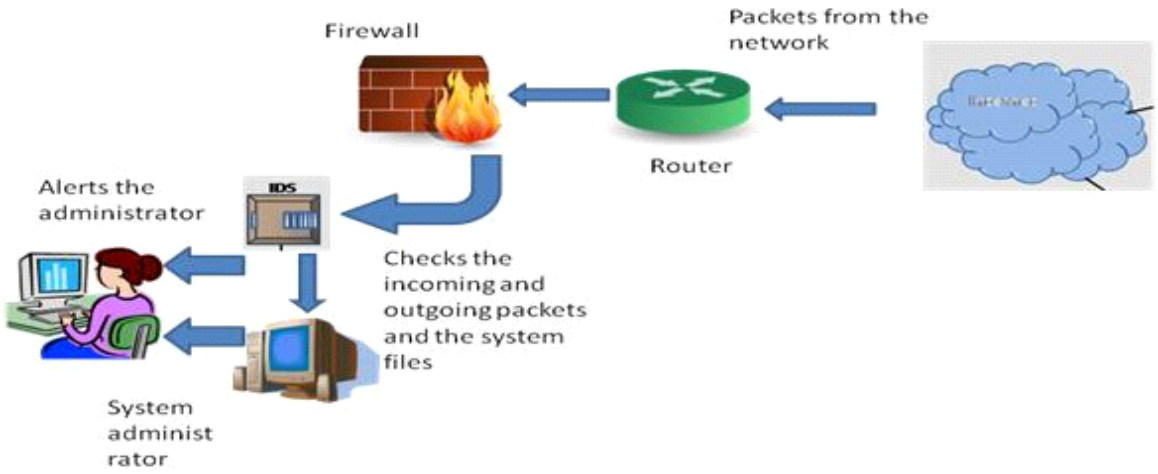
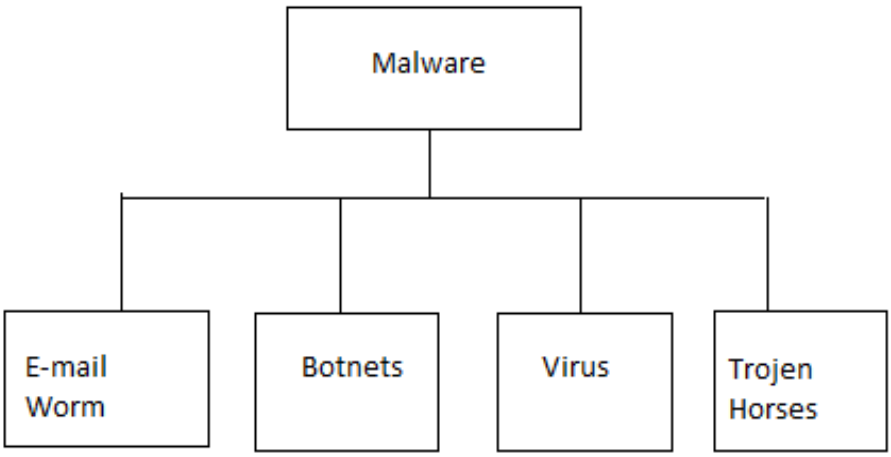


Fig.1: Intrusion Detection System

1.1 Data Mining

Data mining computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database [1] systems. Also known as knowledge discovery in databases, an interdisciplinary subfield of the data mining step might identify many groups in the information and the data is used to obtain more precise results by system. Data mining is the process of posing queries and extracting patterns, often previously unknown from large quantities of data using pattern matching or other reasoning techniques. Data mining have various applications in security including for national security as well as for cyber security.

Extract information is the goal of data mining process to from a data set of database and transforms it into a structure which is understandable for further use.



.

Fig.2: Data Mining Techniques

2. LITERATURE SURVEY

As mentioned previously most of the current web applications suffer from such threats due to bad coding practices and cross site scripting. Spyware poses several risks. The most conspicuous is compromising a user's privacy by transmitting data about that system behaviour. Hence spyware can also detract from the usability and stability of a user's computing environment, and it has the potential to introduce new security vulnerabilities to the infected host. Because spyware is spread wide such vulnerabilities would put lacks of computers at risk. Spyware can be detected through the effects it has on systems: use of system resources resulting in a slowdown of the PC; high level of Internet connection bandwidth consumption; complete loss of the connection or general system instability. They also often change settings of certain applications, such as the browser home page, or insert icons on the desktop. The main [12] consequence of spyware on PCs is the gathering of information, including confidential details, making users feels uneasy about what is happening in their computer behind their backs. In this paper they have done the Anomaly Detection Techniques could be used to detect unusual patterns and behaviors. Link analysis may be used to trace the viruses [11] to the perpetrators. In the paper explained the techniques for detecting and analyzing Malware executables. Computer system's security is threatened by weapons named as malware to accomplish malicious intention of its writers. Many solutions are available to find these threats like AV Scanners, Intrusion Detection System, and Firewalls prediction etc. [10]. These solutions of malware detection classically use signatures of malware to detect their presence in our system.

3. PROPOSED SYSTEM

Using classification profiling of the input code is done so as to differentiate between good and malicious code. Prediction is used based on the previous metrics and classification so that if a code sample does not fit a given skeleton then based on previously generated ones'

we can abstract the given test case to categorize the input. Moreover firewalls can be configured using the given guidelines which will help avoid the threat of malicious code.

4. MALICIOUS CODE DETECTION

Malware mainly includes different computer viruses, ransom ware, Trojan horses, worms, root kits[10], key loggers, adware, dialers, spyware, rogue security software's and some other malicious programs; the majority of active malware threats are normally Trojans or warm rather than viruses. Malware is kind of a virus, worms, Trojans, adware's, spywares root kit, etc. Malware is known as computer pollution, as in the legal rules of several United States. Malware is different from unusable software [10], which is legitimate software but having harmful bugs that were not removed before release. Spyware is any software installed on a computer without the user's knowledge that gathers information about that user for later retrieval by whoever controls it. There are two types of spyware: malware and adware. Malware is any program that gathers personal information from the user's PC. Key loggers, screen capture devices, and Trojans are in this category. Adware is a program designed for showing user advertisements, like homepage hijackers, pop-up windows and search page hijackers. Spyware poses several risks. The most vulnerable is compromising a user's privacy by transmitting information about that user's system behaviour. However, spyware can also distracts from the usability and stability of a user's computing sector and it has the potential to introduce new security vulnerabilities to the infected host. Because spyware is spread wide, such vulnerabilities would put millions of computers at risk.

5. SYSTEM ARCHITECTURE

The system architecture produce the input file of classified header, its functions and miner operation to be carried out over the feature database for to selection and transformation of the desired application to test over the trainer and to set the various testing properties to the trainer to learn the required algorithms for the testing and generated classified results.

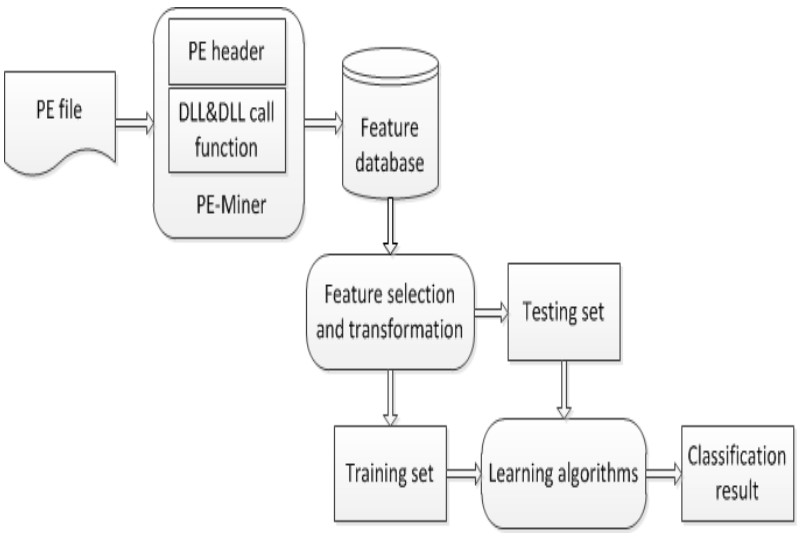


Fig.3: System Architecture

6. SOFTWARE REQUIREMENT SPECIFICATION

6.1 Project Scope

Project proposes using Data Mining method for Detecting new malicious executables. We apply three different algorithms with each one having its own feature extraction Techniques. These three algorithms are as Ripper, Naives Bayes approach, and Multi-Naive Bayes using data mining techniques and the comparison of these algorithms.

6.2 User Classes and Characteristics

- Get Data – Used to collect data from sources
- Dump To Warehouse – the data that is being collected is dumped onto warehouse
- Analyze – This class is defined to contain methods of analysis of the given data
- Show Details – Responsible for display and GUI related activities

6.3 Operating Environment

Data to be mined is collected from web crawlers which are then fed to an algorithm in mat lab. The analysis is done in mat lab itself where in new rules are defined so as to counter the threat.

6.4 Design and Implementation Constraints

Spyware program have a series of traits that makes them difficult to detect and enables them to install themselves on numerous PCs for long period of time.

1. They use almost perfect camouflage systems. They are normally installed on computers along with another kind of application: a P2P client, a hard disk utility.
2. File names don't normally give any clues as to [12] the real nature of the file, and so they often go unnoticed along with other legitimate application files.
3. As they are not viruses, and they don't use a routine [12] that can be associated with them, antivirus programs don't detect them unless; they are designed specifically to do so.

Current virus scanner technology has two parts: a signature-based detector and a heuristic classifier that detects new viruses. The classic signature-based detection algorithm [11] relies on signatures (unique tell-tale strings) of known malicious executable to generate detection models. Signature-based methods create a unique [11] tag for each malicious program so that future examples of it can be correctly classified with a small error rate. These methods do not generalize well to detect new malicious binaries because they are created to give a false positive rate as close to zero as possible. Whenever a detection method generalizes [11] to new instances, the trade-off is for a higher false positive rate. [10]Heuristic classifiers are generated by a group of virus experts to detect new malicious programs. This kind of analysis can be time-consuming [12] and oftentimes still fail to detect new malicious executables.

6.5 Assumptions and Dependencies

It is assumed that the data we gather contains malicious code within itself. During the implementation phase this data is analyzed based on the machine learning and data mining algorithms such as Naive Bayes and using this pattern is found out which we can say as the signature of the malicious code.

7. METHODOLOGY

We propose using data mining methods for detecting new malicious executables. They apply three different algorithms with each one having its own feature extraction technique. The first one is RIPPER algorithm, which [12] they only applied on Portable Executable (PE) format data using the Portable Executable header information extraction technique, so I skip this algorithm. The second algorithm is Naïve Bayes algorithm using strings in the binaries as features. This technique can be easily avoided by encoding or encrypting the strings in a file, so this technique is weak against new malicious [12] code. The third algorithm is Multi-Naïve Bayes algorithm, which uses byte sequences in a file as features. We can detect malicious executable by looking at the frequency analysis of byte code in a file. Multi-Naïve Bayes algorithm is basically a collection of Naïve Bayes algorithms for splitting the data into sets. The dataset consists of 312 benign (non-malicious) sources and 614 malicious sources. The spyware collection is formed by using the virus collection at http://vx.netlux.org and by crawling the internet using [12] a sandboxed operating system and manually collecting spywares. The benign executables are collected from the system files in Windows XP operating system and from programs of a stereotype user. Byte sequences are extracted using the hex dump tool in Linux. For each file in the dataset, using this tool a hex dump file is formed. When the algorithm is run, user should specify a “window size”. Naïve Bayes algorithm can be specified to use how many sequences of byte data. Default window size is one, which means algorithm will look at 4 hexadecimals (2 bytes of data) for frequency analysis. I run the algorithm for a window size of [12] 2 and 4 separately using 5-fold cross validation technique.

8. CONCLUSION

Data mining-dependent malicious code detectors have been very successful in detecting malicious code such as viruses and worms. Therefore we successfully implemented Data mining to provide solutions such as intrusion detection and auditing. Data mining is also being applied for intrusion detection and auditing. Other applications are also successfully implemented data mining for malicious code detection such as worm detection and managing firewall policies. Secondly concluded the various types of threats to national security and describe data mining techniques for handling such threats. Threats include non real-time threats and real-time threats. Also implemented resulted Data mining applied for credit card fraud detection and biometrics related applications. Progress has been made on topics such as stream data mining; there is still a lot of work to be done here and concluding we have discussed the consequences to privacy for Data mining. It is expected that this procedure will lead to the development of better algorithms for identifying the root kit that has infected a system.

ACKNOWLEDGMENT

We would like to sincerely thank Prof. Vinod Wadne our guide for her support and encouragement.

REFERENCES

[1]. Data mining for malicious code detection and security application http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=52487&queryTxt%3DData+Mining+for+Malicious+Code+Detection+And+Security+Applications

[2]. International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4, July 2013, A static malware detection system using data mining methods

[3]. I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Ed. Morgan Kaufmann, 2005.

[4]. Data mining for malicious code detection and security application http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6061180 12-14 Sept. 2011Intelligence and Security Informatics Conference (EISIC), 2011 European

[5]. J. Z. Kolter and m. A. Maloof, "learning to detect malicious executables in the wild," in Proceedings of The acm symp. On knowledge discovery and data mining (kdd), pp. 470-478, August 2004.

[6]. Guillermo Suarez-Tangle, "Evolution, Detection and Analysis of Malware for Smart Devices" IEEE Communications surveys & tutorials, accepted for publication, pp.1-27, 2013.

[7]. Kirti Mathur, "A Survey on Techniques in Detection and Analyzing Malware Executables" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

[8]. Malicious PDF Document Detection Based on Feature Extraction and Entropy by Himanshu Pareek, Centre for Development of Advanced Computing, Hyderabad, India. International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 2, No 5, October 2013.

[9]. Classification of Malware based on Data Mining Approach, IJSRD - International Journal for Scientific Research & Development| Vol. 1, Issue 2, 2013 | ISSN (online): 2321-0613.

*[10]. Intrusion Detection System using Data Mining, Volume 2, Issue 6, June 2014. International Journal of Advance Research in **C**omputer Science and Management Studies. ISSN: 2321-7782 (Online).*

[11]. Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM, International Joint Conferences on 10.1109/WI-IAT.2009.372 Sep-2009.