IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Cyber Security Approach using Data Mining for Malicious Code Detection

Ms. Poonam Bhagwandas Godhwani[1], Ms. Jayshri Arjun Patil[2]

Teaching Assistant, Department of Computer Science & Technology, UTU-Bardoli (Gujarat) [1, 2]

Abstract: A serious security threat today is malicious executables, especially new, unseen malicious executables often arriving as email attachments. These new malicious executables are created at the rate of thousands every year and pose a serious security threat. Current anti-virus systems attempt to detect these new malicious programs with heuristics generated by hand. This approach is costly and oftentimes ineffective. In this paper, we present a data-mining framework that detects new, previously unseen malicious executables accurately and automatically. The data-mining framework automatically found patterns in our data set and used these patterns to detect a set of new malicious binaries. Comparing our detection methods with a traditional signature based method, our method more than doubles the current detection rates for new malicious executables.

Keywords: serious security threat, ineffective, accurately and automatically

INTRODUCTION

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 has many applications in security including for national security as well as for cyber security. A serious security threat today is malicious executables, especially new, unseen malicious executables often arriving as email attachments. These new malicious executables are created at the rate of thousands every year and pose a serious security threat. Malicious code continues to evolve and create new challenges for organizations seeking to protect themselves.

The task of actual data mining is automatic or semi-automatic analysis of huge amount of data to extract previously unknown interesting patterns, unusual record and dependencies. It usually involves in using database techniques such as spatial indices. These patterns can be seen as details of the input data, and may be used for further analyzing e.g. in machine learning and predictive analytics. For example, the data mining step might identify many groups in the information and the data is used to obtain more precise results by system. The data collection, data preparation, result interpretation and reporting are not the part of the data mining step, but belongs to the overall knowledge discovery process as further step.

Data mining has many applications in security including in national security (e.g., surveillance) as well as in cyber security (e.g., virus detection).

Malicious code is a term used to describe any code in any part of a software system or script that is intended to cause undesired effects, security breaches or damage to a system. Malicious code describes a broad category of system security terms that includes attack scripts, viruses, worms, and Trojan horses, backdoor and malicious active content. Malware mainly includes different computer viruses, ransom ware, Trojan horses, worms, root kits, key loggers, adware, dialers, spyware, rogue security softwares and some other malicious programs; the majority of active malware threats are normally Trojans or warm rather than viruses. Malware is known as computer pollution, as in the legal rules of several United States. Malware is different from unusable software, which is legitimate software but having harmful bugs that were not removed before release. However, some malwares are masked as genuine software, and may come from any website in the form of useful program which has the harmful malware included in it with additional tracking software that gathers information.

Malicious software (malware) is any software through which the creator of malware can take the full and partially to full control of your computer whatever the developer wants. Malware is kind of a virus, worms, Trojans, adware's, spywares root kit, etc [10]. Spyware is a kind of malware installed on computers which takes information about users without their knowledge. Artificial Intelligence was established during a conference. The technology gets so wide and evolutes many other branches of engineering field like electronic, robotic etc. This mainly led to useful for complex and smart machinery. By the evolution of malware detection system and Artificial Intelligence (AI), as a latest technology, Artificial Intelligence (AI) has been implemented in anti-viruses engines. There are several Artificial Intelligence approaches that implemented in spyware detection systems such as ANN, Heuristic Technology and Data Mining Technique.

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

DOI 10.17148/IJARCCE.2016.56157

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.
Current anti-virus systems attempt to detect these new malicious programs with heuristics generated by hand. This approach is costly and oftentimes ineffective.

How Data Mining is useful in detecting malicious code?
In this system we are using data mining methods, we were planning a system that will automatically design and build a scanner that accurately detects malicious executables before they get a chance to run on the system.

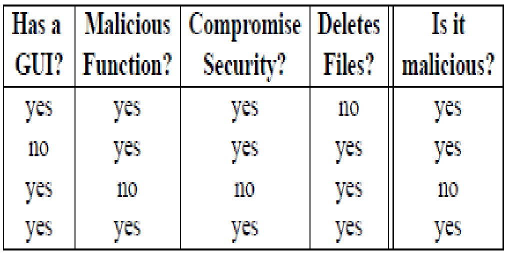
Data mining methods detect patterns in large amounts of data, such as byte code, and use these patterns to detect future objects in similar data in the database. Our framework uses classifiers to detect new malicious executables. A classifier is a set of rules, or detection model which is generated by the data mining algorithm that was trained on a given set of training data.

One of the primary problems faced in today's world by the virus community is to find methods for detecting new malicious programs that have not yet been analyzed before Many malicious programs are created every day in the world by hackers and most cannot be accurately detected until proper signatures have been generated for them. During this time period, systems protected by signature-based algorithms are vulnerable to attacks on such malicious codes.

ALGORITHM AND 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. RIPPER is a rule-based learner that builds a set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules.

An inductive algorithm learns what a malicious executable is given a set of training examples. The four features
seen in Table 2 are:
1. "Does it have a GUI?"
2. "Does it perform a malicious function?"
3. "Does it compromise system security?"
4. "Does it delete files?"
and finally the class question "Is it malicious?"



Has a GUI?	Malicious Function?	Compromise Security?	Deletes Files?	Is it malicious?
yes	yes	yes	no	yes
no	yes	yes	yes	yes
yes	no	no	yes	no
yes	yes	yes	yes	yes

Table 2: Example Inductive Training Set. Intuitively all malicious executables share the second and third feature, "yes" and "yes" respectively.

The defining property of any inductive learner is that no apriori assumptions have been made regarding the final concept. The inductive learning algorithm makes as its primary assumption that the data trained over is similar in some way to the unseen data.

CONCLUSION

Data mining-dependent malicious code detectors have been very successful in detecting malicious code such as viruses and worms. There are many techniques that has been developed till now that can dynamically adapt to new detection strategies and continued to monitor the adversary. There is a need for a technique in which detection of malicious patterns in executable code sequences can be done more efficiently. Moreover with a larger data set, we can evaluate data description method on many types of malicious executables like macro and Visual Basic script.. We can extend our algorithms to utilize byte sequences in future .There is a need to implement the method on the interconnected computers for evaluating the performance in terms of time, space and accuracy in real world environments so that we can detect the attacks in larger data sets efficiently.

REFERENCES

[1] William Cohen. Learning Trees and Rules with Set-Valued Features. American Association for Artificial Intelligence (AAAI), 1996.
[2] Thuraisingham, B., "Web Data Mining Technologies and Their Applications in Business Intelligence and Counterterrorism", CRC Press, FL, 2003.
[3] M. G. Schultz, E. Eskin, E. Zadok and S. J. Stolfo, "Data Mining Methods for Detection of New Malicious Executables", Proceedings of the 2001 IEEE Symposium on Security and Privacy, IEEE Computer Society.
[4] Bhavani Thuraisingham, Data Mining for Security Applications, IEEE/IFIP International Conference on Embedded and Ubiquitous Computing,2008 .
[5] Dr.R.Geetha Ramani, Suresh Kumar.S , Shomona Gracia Jacob"Rootkit (Malicious Code) Prediction through Data Mining Methods and Techniques" , 978-1-4799-1597-2/13/$31.00 ©2013 IEEE.
[6] Guillermo Suarez-Tangle, "Evolution, Detection and Analysis of Malware for Smart Devices" IEEE communications surveys & tutorials, accepted for publication, pp.1-27, 2013.

DOI 10.17148/IJARCCE.2016.56157