# Malicious Code Detection through Data Mining Techniques

Ms. Milan Jain
Research Scholar
Department of CSE
Chandigarh Engineering College
Mohali, Punjab, India
milan.jain2@gmail.com

Ms. Punam Bajaj
Assistant Professor
Department of CSE
Chandigarh Engineering College
Mohali, Punjab, India
Cecm.cse.punb@gmail.com

**Abstract- Nowadays computer systems and communication infrastructures are likely to be influenced by different types of attacks so there is need to put further efforts for improving the software trust. Therefore, there will be increase in necessity in the coming time, as the number of software developers and applications will likely grow very significantly. As important advances have been already made on malware executables detection in personal computers in the previous decades which we have reviewed in previous works. However there is more need to adopt some better techniques which can ensure the malware code detection efficiently by testing method over a large set of malicious executables. This paper explores the application of data mining methods to predict rootkits based on the attributes extracted from the information contained in the log files. The rootkit records were categorized as Inline and Other based on the attribute values. In this paper, we proposed three algorithms named as RIPPER, Naives Bayes approach, and Multi-Naïve Bayes using data mining techniques and the comparison of these algorithms.**

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

## I. INTRODUCTION

Rootkits is 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. Installing a rootkit is usually the first basic step that an attacker will do after gaining access to a system, as this thing is going to tell us that the attack will remain undetected. Therefore attacker can then further proceed to capture the personal data such as bank account details, passwords, and credit card numbers [3]. To ensure the integrity of computer networks in relation with security and the privacy is the important documents of nation is a great concern. Defense and security with networks, intellectual property proprietary research, and data based market mechanisms that depends upon unimpeded and undistorted access can fully settled by malicious intrusions. We need to find an optimum way to save these types of systems. To perform this we require different techniques to detect security breaches. Data mining support many applications in security that coincides in national security as well as in cyber security. The different types of threats related to a country national security include attacks on buildings and demolish critical infrastructures such as telecommunication network systems and power system grids [1]. 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. 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. The main difference between such type of “smart” devices and “non-smart” equipments is easily interconnect with third-party applications through online markets. Smart devices popularity is increasing very widely because these are formally related to rise of cloud-computing approach provides complementary storage and computing services [5].

## II. DATA MINING

Data mining also known as knowledge discovery in databases, an interdisciplinary subfield of 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 systems. 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. Aside from the analysis step method, it has database and data management aspects, data pre-

processing, model and inference considerations, complexity considerations, interestingness metrics, visualization, post-processing of discovered structures, and online updating.

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.

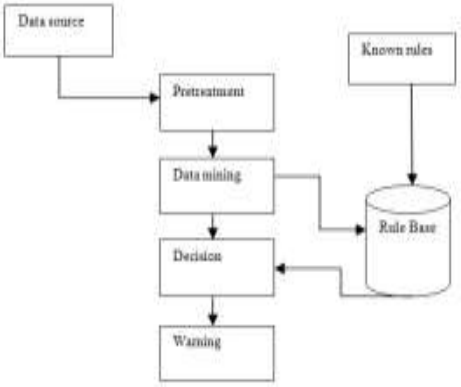


Fig1.Data Mining Techniques

## III. MALICIOUS CODE DETECTION

Malware the malicious software is used to gather sensitive information, disrupt computer operation or to have access to secure computer systems. It can be appear in the form of coding, scripts, active contents and other software. Malware is the term used to refer a variety of forms of intrusive software [8].

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 viruses, worms, Trojans, adwares, 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 get 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. Heuristic-based Detection performs well opposite to known Spyware but is not yet proven very successful for detection new spyware. The growth in high-speed Internet connections in today's life increases malware to propagate and effects hosts very fast and easily, so it is necessary to identify and eliminate malwares in a manner. Anti-virus vendor is also facing a number of such suspicious files every day [9]. These files are received from various sources including honey pots, third party transfers and files send by customers either automatically or manually. The huge amount of files is not being able to inspect efficiently and effectively. The main aim of this study is to be able to check out not known malicious files from the files arriving into the internet every day, and to remove these malicious files [10].

## IV. LITERATURE SURVEY

In the paper [8] explained different methods of detecting a malicious executables. These malicious executables are formed at the huge rate every year and create a serious security threat. The anti-virus systems attempt to detect these malware programs with heuristics created by hand. This method is costly and sometime less-effective. In this paper, present a data-mining platform that detects new malicious files effectively and automatically. The data-mining platform automatically detects patterns in their data set and used these detected patterns to identify a set of new malicious executables. Compare these detection methods with a classical signature based method, the new method provides doubles the current detection rates for unseen malicious files.

In the paper [6] explained a model checking method for detecting malicious code. In this paper, author presents a soft method to detect malicious code sets in executables files by using model checking. While model checking was developed to check the correctness of system against specifications, author commented that it grants equally well to the identification of malicious code patterns. In the end, they introduced the specification language Computation Trees Predicate Logics which is extending the well-known logics CTL and gave description about an efficient model checking approach their practical experiments demonstrate that they are able to detect a large number of worm variants with a single specification.

In the paper [1] explained various data mining techniques for security application. These requisition include but are not limited to malicious executables detection by mining it binary executables, anomaly detecting and data stream mining process. They summarize their acquirement and present works at the University of Texas at Dalla on intrusions detection and cyber-security research.

In the paper [7] 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. Various solutions are available to detect these threats like AV Scanners, Intrusion Detection System, and Firewalls etc. These solutions of malware detection traditionally use signatures of malware to detect their presence in our system. But these methods are also evaded due to some obfuscation techniques employed by malware authors. This survey paper highlights the existing detection and analysis methodologies used for these obfuscated malicious code.

In the paper [5] ,showed malware in current smart devices that equipped with powerful sensing, computing and networking capabilities have proliferated lately, range from famous smart android phones and tablets to Internet devices, smart TVs, and others that will soon appear. One main feature of devices is that they have ability to incorporate third-party applications from markets. This has very strong security features and secrecy problems to user and infrastructure operator, specifically via software of malicious nature that got access to the service given by the devices and gather the sensory data and personal data. Malware in latest smart devices – Smart phones and tablets– has got fame in the previous few years, in some cases supported by best techniques designed to provide better security architecture presently in use by these devices. As important advances have been made on malware detection in computers in the last decades it is still a challenging problem.

In the paper [9] implemented artificial Intelligence in anti-virus engines. Malicious software is the software which gives partial to full control of your computer to do whatever the malware creator wants. Malware can be defined as a viruses, worms, Trojans, adwares, spywares and root kits. Spyware is a class of malware which is installed on computer that is able to collect information regarding clients without having knowledge. In 1956, the purpose of establishment of Artificial Intelligence(AI) Dart muth College during a conference. Artificial Intelligence has been implemented in anti-virus engines. AI has many approaches that implemented in spyware detection systems such as Artificial Network, Heuristic Technologies and Data Mining Techniques. In this work, they focused on DM-based malicious code detectors by using Breadth-First Search approach for knowing work well for detection virus and software. BFS is the method for searching in a tree when search is very limited to essentially two operations (a) visit and inspect a node of a tree; (b) gain access to visit the nodes that are neighbor to currently visited node.

## V. PROPOSED METHODOLOGY

Data mining is the process of using information from old data to examine the output of a specific situation that may come. Data mining worked to identify data stored in data warehouses that are being used to store that data that has been analyzed. The specific data can come from all businesses, from the production house to the management. In this work, we are needed to make the system on a network of computers for evaluating the performance to make it more efficient in terms of time and precise in real world environment. It is required to make the learning algorithms more efficient in time and space. Presently, the Naive Bayes methods have to run on a computer with one gigabyte of RAM. Finally we need to plan testing this method over a large set of malicious codes. In this work, we will be using three algorithms presented in this paper and we propose three data mining algorithms to produce new classifiers with separate features: RIPPER, Naive Bayes, and a Multi-Classifier system and the comparison between these three methods. The data mining methodology formulated

for rootkit prediction is diagrammatically presented in Fig.2. It comprises of rootkit data collection, data pre-processing, classification and performance evaluation phases.

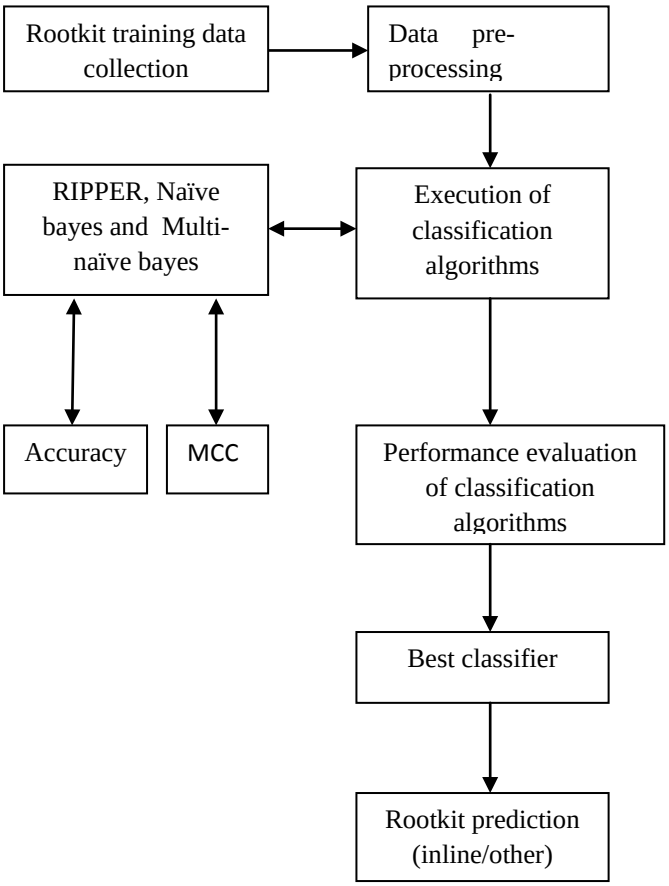


Fig2. Data mining methodology for root kit prediction

*A. RIPPER*

The first algorithm RIPPER is an inductive rule learner [8]. This approach developed a detection model Consists of resource rules that was developed to detect examples of malicious executables. This algorithm is using a LibBFD data as characteristics. RIPPER is a rule-based learning approach that is building a set of rules that is able to determine the classes while reducing the ambiguities. The ambiguity is given by the training examples of unclassified by the rules.

B. *Naive-Bayes*

The next classifiers we are describing is a Naive Bayes classifier. The naive Bayes classifier calculates the likelihood that a program is having malevolent code given the features that are present in the program [8]. This approach used both string and byte sequences data for computing a probability of a binary's malicious code having some features. The main assumption in this algorithm is that the binaries contain same features such as signatures and machine instructions.

*C. Multi-Naive Bayes*

The next data mining algorithm is Multi-Naïve Bayes. This algorithm was importantly a collection of Naïve Bayes algorithms that supported on whole categorization for an example. In Naive Bayes algorithm, it is able to classify the examples in the test set of malicious executables program and this counted as a choice [8]. The votes were combined by the Multi- Naive Bayes algorithm to outcome a final classification for all the Naive Bayes. This method was needed as it is using a machine with 1GB of RAM, the size of the binary data was very big to get into memory. The Naive Bayes algorithm required a table chart of all strings or bytes to evaluate its possibilities. In every classifier, there is a rule set. The divination of the Multi-Naive Bayes algorithm is the multiplication of all the predictions of the Naive Bayes classifiers.

By evaluating these three algorithms, we will find out the best value based on the results of these algorithms named as RIPPER, Naive Bayes and Multi classifier system. Data mining techniques perform better than traditional techniques such as signature-base detection and Heuristic-based detection. Data mining has six

common classes named as Anomaly detection (Outlier/change/deviation detection), Searches for relationships between variables, Clustering ,Classification ,Regression, Summarization.

## VI. 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. This paper explores the application of data mining methods to predict rootkits based on the attributes extracted from the information contained in the log files. It is expected that this procedure will lead to the development of better algorithms for identifying the rootkit that has infected a system.

## REFERENCES


[1] Bhavani Thuraisingham, Data Mining for Security Applications, IEEE/IFIP International Conference on Embedded and Ubiquitous Computing,2008 .

[2] D.Michie, D.J.Spiegelhalter, and C.C.TaylorD. Machine learning of rules and trees. In Machine Learning, Neural and Statistical Classification. Ellis Horwood, 1994.

[3] 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.

[4] E. Chin, A. P. Felt, V. Sekar, and D. Wagner, "Measuring user confidence in smartphone security and privacy," in Symp. on Usable Privacy and Security. Washington: Advancing Science, Serving Society, March 2012.

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

[6] Johannes Kinder, "Detecting Malicious Code by Model Checking" *pure.rhul.ac.uk/portal/files/17566588/mcodedimva05.pdf.*

[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] 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.

[9] Parisa Bahraminikoo "Utilization Data Mining to Detect Spyware", IOSR Journal of Computer Engineering (IOSRJCE),Volume 4, Issue 3, pp.01-04,2012.

[10] Robert Moskovitch "Detecting unknown malicious code by applying classification techniques on OpCode patterns" Springer-Verlag "http://link.springer.com/article/10.1186%2F2190-8532-1-1" 2012.

[11] S. Larner, "Smartphones and tablets in the hospital environment," British J. of Healthcare Management, vol. 18, no. 8, pp. 404–405, 2012.

[12] Wildlist Organization. Virus descriptions of viruses in the wild. Online publication, 2000. http://www.fsecure. com/virus-info/wild.html.

[13] X. Wei, N. C. Valler, B. Prakash, I. Neamtiu, M. Faloutsos, and C. Faloutsos, "Competing memes propagation on networks: A network science perspective," IEEE J. Sel. Areas Commun., vol. 31, no. 6, pp 1049–1060, 2013.

[14] Y. Lee, Y. Ju, C. Min, J. Yu, and J. Song, "Mobicon: Mobile context monitoring platform: Incorporating context-awareness to smartphonecentric personal sensor networks," in 9th Annu. IEEE Commun. Society Conf. on Sensor, Mesh and Ad Hoc Commun. and Netw. (SECON 2012), 2012, pp. 109–111.