International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
International Manuscript ID : 22296166V7I1101602
Volume 7 Issue 1 January 2016

Malware Detection using Data Mining Techniques: A Review

Prasenjit Das
Dept. of Computer Science & Engineering,
Chitkara University , Himachal Pradesh
prasenjit.das@chitkarauniversity.edu.in

Sumeet Dua
Dept. of Computer Science & Engineering
Chitkara University , Himachal Pradesh
sumeetdua_usa@gmail.com

Abstract

Malware detection techniques are rendered ineffective because of the large number of variants being generated from time to time. The large no. of families that attack the systems day in and day out have created havoc on the systems that are either standalone or connected on the web. Data mining techniques which capture the behaviour of the malware have been proved effective in tracking the malware attacks.

Keyword: Data Mining, Malware, Roolkit, Malware Family

1. Introduction

Information that is available on the standalone as well as network systems is constantly under the threat of being attacked by software trying to malign the working of the system or steal the data.Malware is a collective term which is used to describe any software that enters, disturbs operations, and gathers information from the system without the knowledge of the user/users. The term is coined as a combination of two terms 'malicious' and 'software'.

Malwares are disguised in form of non malicious files which comes into effect when the file is being used. It can attach itself as an executable file, a script running on an internet browser, or an activeX^{TM} control.Malwares either infect a host file or system area, or they simply modify a reference to such objects to take control and then multiply again to form new generations(Szor, 2005). Once a malware enters into a system, malware is capable of checking the surfing habits of the user, spying on the logging password by observing the key strokes, sneaking into read/unread mails, hacking the web browser and connecting web pages containing user details to unauthorized websites and variety of other malicious activities.

In the recent past widely publicized attacks suggest that the malwares are becoming a critical problem for industry, government, and individuals. Malware attacks have occurred across the globe, across industries have hampered the working of the same. Redirecting of websites of

major news agencies like BBC™, New York Times™(Hern, 2016), stealing millions from banks("New GozNym banking malware steals millions in just days," n.d.), modifying records of patients in hospitals("Malware Attacks On Hospitals Put Patients At Risk," n.d.) are very few instances of the damage being done by malwares. Some of the signs that let the user know that a malware has entered into their personal computers are:

- Start seeing an excessive amount of pop-up ads.
- PC's operating system slows down significantly.
- The amount of spam you receive in your email increases.
- Email account may send out messages to your contact list that you did not send.
- The home page you have set in your browser is altered.
- While trying to access a web page in your favorites list, another web page appears that contains advertising or content that encourages you to enter your personal information.
- Computer completely crashes.
- Unable to access your antivirus program to remove the malware.

1.1 Malware Families

In normal routine, most of the malwares are referred to as a virus. In case of systems showing any of the above mentioned symptoms, it is referred to as a virus attack. But not all 'virus attacks' are caused by viruses. Based on the way the malware enters into the system, way they execute and damage that ach malware can cause, malwares are categorized into different families/types. Categorization into malware families is also done on basis of following parameters:

i) Amount of code sharing
ii) Language used to develop the malware
iii) Malware using same Libraries
iv) Way the packing of the malware is done

As per Microsoft ™, there exists close to 200 families of malwares. Following is a detailed explanation of major malware families

Computer Virus:

Computer virus is the type of malware that is and commonly used to attack a system. It is one of the most widely discussed malware when it comes to the security of both standalone and network systems. Here a virus manifests as a computer program that is designed to provide unauthorized control of the host system so as to replicate, delete, modify the system's data or to consume the systems resources thus degrading the operation of the system altogether.

Trojan horse:

A Trojan horse or simply called as Trojans is a type of malicious program that disguises itself as something that is legitimate or useful.The main purpose of a Trojan is to gain the trust of

the user from the front end, so that it gets the permission to be installed. However it is designed to grant unauthorized control of the system to the hacker.

Worms:

Worms are standalone computer programs with a malicious intent that spread from one computer to another. Unlike viruses, worms have the ability to operate independently and hence do not attach themselves to another program.

Spyware:

Spyware is a type of malicious software that can collect information about the activities of the target system without the knowledge of its users. Spywares such as keyloggers are often installed by the owner or administrator of the computer in order to monitor the activities of the users. This can be a parent trying to monitor his child, a company owner trying to monitor his employee or someone trying to spy on his/her spouse.

Adware:

Adware is a software program that automatically renders advertisements to the users without their consent. Most common examples are pop-ups, pop-unders and other annoying banner ads. The prime reason behind the design of adware is to generate revenue for its author.

2. Data Mining

Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, andpotentially useful information from data"(Frawley, Piatetsky-shapiro, & Matheus, 1992) and "the science of extracting useful information from large data sets or databases" (Hand, Mannila, & Smyth, 2001). Data mining tool relies on Knowledge Discovery in Data, KDD for short. It is the process of extracting the unknown from a raw data. Data in its raw form is collection of elements out of which no fruitful information can be gleaned. There are number of steps involved (explained below) in extracting patterns from the data which when interpreted provides unknown information. The extracted information can be predictive or descriptive. Predictive data mining involves predicting values based on the given data. Descriptive data mining involves finding patterns describing the data. Many fields, mostly which are involved in decision making are using the KDD process to convert data into information.

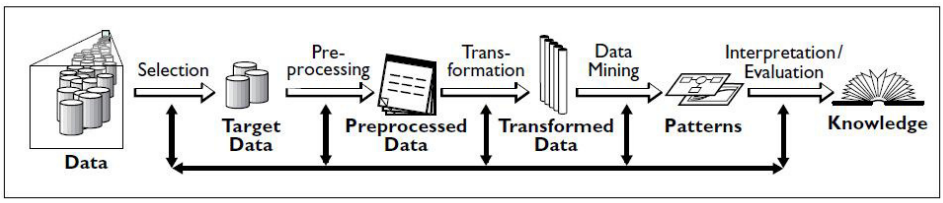

Figure 1: Overview of Steps in KDD process

Source: http://www.ceine.cl/the-kdd-process-for-extracting-useful-knowledge-from-volumes-of-data/

3. Motivation

The failure of traditional methods in detecting malwares resulted in the research work being carried out for detection of malware by tracking the behaviour of the malware. The attributes in a particular binary executable distinguishes the executable as a malware or benign software. Though this approach has been fruitful in detecting the malicious behaviour of the malware yet the malware being introduced in the computer fraternity is increasing each day. The counter measures available at any instance to deal with malware simply outnumber the no. of new families or their variants that are generated each day. Most of the malwares that are generated are not always new code. It is just a variant of an already existing code manipulated by the coder to perform a new malicious activity. This makes us believe that there does exists a relation between a different instances of malware (with same family or two different families) in terms of the features/attributes each potential binary executable has. This belief is a motivation for us to explore the dynamic behaviour of a binary executable to detect the malicious activity in a binary executable.

4. Related Work

Traditional malware detection techniques have been rendered ineffective as the unknown malwares (Zero Day Attack) do not have a known signature tell a tale. Using Data mining as a tool for malware detection, 3 types to malware detection techniques have been developed.

i) Anomaly based detection involves creating a normal profile of system or program and check for any change/deviation from the profile
ii) Misuse based detection involves creating a malicious profile and checking for its signature in the yet unknown executable
iii) Hybrid based detection does not create any profile but uses both benign and malicious profile to build classifier.

4.1 Traditional Methods

Static and dynamic malware detection techniques have been proposed using data mining as a tool in most of literature of malware detection. Static malware detection has not been able to track or nullify the zero day attacks because of the non availability of signatures of unknown malwares. Static Analysis depends on reverse engineering tools such as disassembler to transform binary codes to assembly codes, comparatively readable by human. Key binary codes, extracted as a signature of the code may include one or a combination of the following: byte n-gram, opcodes n-gram, function calls, PE features, Strings, and Control Flow Graph (Schultz, Eskin, Zadok, & Stolfo, 2001). Christodorescu et al (Christodorescu & Jha, 2003) started early research on static analysis focusing on the specification of the binary codes, while Schultz et al (Schultz et al., 2001) extracted key static features, such as bytes, DLLs, and strings, for malware classification.

4.2 Behavior Analysis

Behavior analysis (also called dynamic analysis) allows an analyst to execute the binary code in a sandbox or a virtual environment, in which malware interaction (file, process, and

network) with the system will be closely monitored and recorded. Kolter et al (Kolter & Maloof, 2006), Rieck et al (Rieck, Holz, Willems, Düssel, & Laskov, 2008) and Firdausi et al(Firdausi, lim, Erwin, & Nugroho, 2010) extracted behavior features from running in a sandbox and classify the result using various machine learning algorithm.

A framework which integrates the static evasion detection with static and behavior features for fast and efficient malware analysis has been presented in (Lim & Ramli, 2015). The hindrance in classification of malware using the traditional method of signature detection has been addressed by using structured control flow graphs (Cesare & Xiang, 2010b). The paper provides a classification model for unpacking the binary and classifies it as malicious or non malicious. It his assumed that Malware's control flow information provides static analysis a characteristic that is identifiable across strains of malware variants. This characteristic is shared across malware families because malwares are created often using the same code. This reuse of code can be identified through isomorphic and similar flow graphs.

4.3 Graph Method

Call graph (Cesare & Xiang, 2010a) have been used as a malware detection method. Detection of malware has been proposed using matching of call graph for malware. Call graph is a directed graph which represents the relation between various subroutines in a computer program. The paper assumes that most of the polymorphic malwares share the same code of history. The obfuscation done to go undetected in presence of anti viruses is done in the packing part of the malware. Sharing the same code history means that the execution flow of the malwares are similar to one another and if the similarity of the graphs (in flow control) is detected between a known malware and unknown binary then the binary can be branded as a malware. A graph edit distance is measured which is used to check for the similarity between two programs and if the similarity between a binary executable and a known malware is approximately same then the binary executable can be termed as infected one. The paper proposes a solution based on control flow graphs in which the similarity between the known malware and unknown variant is calculated using the dice coefficient.

4.4 Classification

Associative Classification (Komashinskiy & Kotenko, 2010) using the post processing techniques have been proposed and compared with other malware detection systems. The paper assumes that by taking into consideration the sequence of bytes (static detection method) can be used to detect features which are at particular position in binary executable based on the position of the features the binary exe can be classified into a malware. Post processing techniques like rule pruning has been used on the known signatures of viruses and a classifier has been built. Naïve Based classification technique has been used. The classifier has been able to detect malwares better than the systems compared in the said literature. The proposed methodology takes into consideration the signatures of known malwares only which can hinder the detection of malwares in case of zero day attack. Similar work has been proposed in the papers (Dai, Guha, & Lee, 2009), (Schultz et al., 2001),(Zhang, Yin, & Hao, 2007).

4.5 Feature Extraction

Extracting the features in a Binary executable that triggers the malicious activity has been proposed (Kolbitsch, Holz, Kruegel, & Kirda, 2010) in form of a gadget which automatically checks for the log file in the system. From the log file the gadget is able to identify the features that actually triggered the malicious activity and hence curb the same. The slicing algorithm proposed in the paper is useful when only one thread is active. But in case of multi threaded system which normally most of the operating systems are, the gadget's accuracy decreases. Dimensionality reduction has been a major challenge in case of Data mining. Decrease in the detection rate of malware can be attributed to the existence of irrelevant and redundant features. Correlation based feature selection method has been used in (Jiang, Zhao, & Huang, 2011) to remove redundant and irrelevant features. Corresponding features of different classes of data has been used in feature selection to reduce the dimensionality. Another method of dimensionality reduction has been used by Mohamad Masud et. al in (Siddequi, Wang, & Lee, 2008). The paper has used Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. The dimensionality reduction has been compared by taking into account feature selection and then using decision tree J48. The data set used in the experiment of was not complex hence the best results were obtained using the decision tree.

5. Conclusion

Malware have been a real challenge for the computer systems across the world. And with the connectivity amongst the systems no pc/laptop is now a standalone. This makes the threat of malware all the more powerful. The various families of malware discussed in the present work keep the anti-malwares on their toes. Traditional techniques used to detect malware were based on the signatures which are ineffective because any new variant of malware is not detectable as the signatures are not available. The methods to track malware based on their behaviour are more effective in the detection and prevention of the attacks. Tracking the behaviour based on the function and system calls is able to detect the attack with accuracy.

6. References

Cesare, S., & Xiang, Y. (2010a). A fast flowgraph based classification system for packed and polymorphic malware on the endhost. In *Proceedings - International Conference on Advanced Information Networking and Applications, AINA* (pp. 721–728). https://doi.org/10.1109/AINA.2010.121

Cesare, S., & Xiang, Y. (2010b). Classification of malware using structured control flow. *Conferences in Research and Practice in Information Technology Series*, *107*, 61–70.

Christodorescu, M., & Jha, S. (2003). Static analysis of executables to detect malicious patterns. *SSYM'03 Proceedings of the 12th Conference on USENIX Security Symposium*, *12*, 12–12. https://doi.org/10.1109/MSP.2010.92

Dai, J., Guha, R., & Lee, J. (2009). Efficient virus detection using dynamic instruction sequences. *Journal of Computers*, *4*(5), 405–414. https://doi.org/10.4304/jcp.4.5.405-414

Firdausi, I., lim, C., Erwin, A., & Nugroho, A. S. (2010). Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection. *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 201–203. https://doi.org/10.1109/ACT.2010.33

Frawley, W. J., Piatetsky-shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases : An Overview. *AI Magazine*, *13*(3), 57–70. https://doi.org/10.1609/aimag.v13i3.1011

Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of Data Mining (Adaptive Computation and Machine Learning). *Journal of the American Statistical Association*. https://doi.org/10.1198/jasa.2003.s257

Hern, A. (2016). Major sites including New York times and BBC hit by 'ransomware' malvertising.

Jiang, Q., Zhao, X., & Huang, K. (2011). A feature selection method for malware detection. *Information and Automation (ICIA), 2011 IEEE International Conference On*. https://doi.org/10.1109/ICINFA.2011.5949122

Kolbitsch, C., Holz, T., Kruegel, C., & Kirda, E. (2010). Inspector gadget: Automated extraction of proprietary gadgets from malware binaries. In *Proceedings - IEEE Symposium on Security and Privacy* (pp. 29–44). https://doi.org/10.1109/SP.2010.10

Kolter, J. Z., & Maloof, M. a. (2006). Learning to Detect and Classify Malicious Executables in the Wild. *Journal of Machine Learning Research*, *7*, 2721–2744. https://doi.org/10.1002/asi.20427

Komashinskiy, D., & Kotenko, I. (2010). Malware Detection by Data Mining Techniques Based on Positionally Dependent Features. *Parallel, Distributed and Network-Based Processing (PDP), 2010 18th Euromicro International Conference On*, 617–623. https://doi.org/10.1109/PDP.2010.30

Lim, C., & Ramli, K. (2015). Mal-ONE: A unified framework for fast and efficient malware detection. In *Proceedings of 2014 2nd International Conference on Technology, Informatics, Management, Engineering and Environment, TIME-E 2014* (pp. 1–6). https://doi.org/10.1109/TIME-E.2014.7011581

Malware Attacks On Hospitals Put Patients At Risk. (n.d.). Retrieved from http://www.npr.org/sections/alltechconsidered/2016/04/01/472693703/malware-attacks-on-hospitals-put-patients-at-risk

New GozNym banking malware steals millions in just days. (n.d.). Retrieved from http://www.scmagazine.com/new-goznym-banking-malware-steals-millions-in-just-days/article/489933/

Rieck, K., Holz, T., Willems, C., Düssel, P., & Laskov, P. (2008). Learning and classification of malware behavior. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5137 LNCS, pp. 108–125). https://doi.org/10.1007/978-3-540-70542-0_6

Schultz, M. G., Eskin, E., Zadok, E., & Stolfo, S. J. (2001). Data mining methods for detection of new malicious executables. In *Proceedings. 2001 IEEE Symposium on Security and Privacy, 2001. S&P 2001.* (pp. 38–49). https://doi.org/10.1109/SECPRI.2001.924286

Siddequi, M., Wang, M. C., & Lee, J. (2008). Detecting Internet Worms Using Data Mining Techniques. *Journal of Systemics, Cybernetics & Informatics*, *6*, 48–53. https://doi.org/10.1145/1593105.1593239

Szor, P. (2005). *The Art of Computer Virus Research and Defense*. *Addison-Wesley Professional* (Vol. 43). https://doi.org/10.5860/CHOICE.43-1613

Zhang, B., Yin, J., & Hao, J. (2007). LNCS 4611 - Intelligent Detection Computer Viruses Based on Multiple Classifiers. *LNCS*, *4611*, 1181–1190. https://doi.org/10.1007/978-3-540-73549-6_115