Research Journal of Applied Sciences 11 (10): 942-947, 2016
ISSN: 1815-932X
© Medwell Journals, 2016

Review of Data Mining Techniques for Malicious Detection

Nawfal Turki Obeis and Wesam Bhaya
College of Information Technology, University of Babylon, Babil, Iraq

Abstract: Malicious is the term used to illustrate any code in any part of a software system that is expected to bring about undesired impacts, security breaks or harm to a system. Malicious programming is outlined with a hurtful intent. Recently, malicious detectors attempt to distinguish unwanted codes by checking Application Programming Interface (API) calls using data mining techniques and/or different methods. Matching the API call utilizing data mining strategies can be utilized as a part of malicious detection systems, for example, frequent pattern, clustering, etc. In this study, a review of malicious detection system based on API calls and data mining strategies are taking into account. Each malicious sample is represented as a data of API calls to the data mining techniques. After transforming the sample that input as a simplified data based on data mining techniques, data mining matching calculations are utilized to similarity between the data tested sample and malicious API call tested samples placed in a database. In this study, a review of utilization of various data mining methods for the detection of malicious program.

Key words: Malicious code, malicious detection, API calls, data mining

INTRODUCTION

Malicious program refers to plans that intentionally abuse vulnerabilities in processing frameworks for a destructive reason. Malicious program can be separated between in light of whether the product needs or does not require a host system to work. Another method for classifying malicious program is by recognizing if the product produces duplicates of itself or not (Assad and Deep, 2016).

Malicious program authors regularly utilize different methods to adjust or transform existing malicious into new polymorphic adaptations to evade detection. The accessibility of advanced toolkits has made it less demanding for malware authors to utilize techniques, for example, dead-code addition and register reassignment to perform this change. The malware adjustment or jamming can be classified into polymorphism and transformative nature (You and Kim, 2010; Christodorescu *et al*., 2007).

A malicious detector is a PC program that endeavors to identify and detect malicious utilizing an assortment of methods that incorporate recognizing malicious signature, using heuristic principles and recognizing malicious behavior or activities. Malicious detectors can work locally on the framework that is being protected or give insurance remotely through a PC network (Ravi and Manoharan, 2012).

There are two types of information are required by malicious detectors, in particular, knowledge of the malicious signature or behavior which can be increased through a learning procedure and the system under assessment. Once the two sources of information get to be accessible, the malicious detector utilizes its detection methods to determine if the software is malicious or benign.

These days, a great many examples of possibly harmful executable are submitted consistently for examination to information security organizations which are confronted with the issue of perceiving whether the specimens are malicious or not (Hu *et al*., 2009).

The attackers produce new malicious program from old ones utilizing code obscurity, polymorphism and new conveyance systems, for example, web-assault toolkits which incredibly adds to the noteworthy increment in the quantity of malicious variations being circulated.

Also, there is a tendency among malicious scholars to utilize high-level programming languages to create malicious and processed it into binary afterward which adds more multifaceted nature to the current issue and exhibits the requirement for powerful and effective arrangements (Hu *et al*., 2009).

Literature review: Arnaud and Arnaud provided approximate and exact strategies to calculating the frequent pattern outlier factor without removing any pattern or by extricating a little sample. They propose a calculation that profits the precise FPOF without mining any example. Shockingly, it works in polynomial time based on the size of the pattern.

Corresponding Author: Nawfal Turki Obeis, University of Babylon, College of Information Technology, Babil, Iraq

Aiman proposed the focuses on the detection of outliers in data stream utilizing regular frequent mining method. An outlier measurement is exhibited and a versatile strategy for discovering outliers in stream of information is presented. The results of the empirical studies demonstrated that the proposed methodology is powerful in detecting outliers' data.

Khin and Nyein proposed apply association rule pattern mining approaches for system intrusion recognition framework. In this study, conventional FPgrowth calculation, one of the affiliation calculations is changed and used to mine itemsets from expansive database. The required insights from massive databases are accumulated into a littler data structure (FP-tree). The itemsets produced from FP-tree are utilized as profiles to check malicious in the proposed framework.

Chien-Yi proposed a structure to use anomaly detection and irregular re-examining methods for profiling a client's practices through the frequent patterns of activated system processes. By using the client profiles gained from normal data, our technique can recognize malicious activities and discriminate suspicious activities from various clients.

MATERIALS AND METHODS

Types of malicious detection systems

Behavior/statistical detection: In behavior-based detection, the behavior of normal information has been put in the library. In the event that there is any action which is not happened beforehand, then the report of that action is sent to the system organization. Behavior based detection system can distinguish the assaults which are already obscure through statistical analysis. It is additionally called behavior based detection method as it recognizes the normal and abnormal behavior of client (Shah and Singh, 2015; Egele *et al*., 2012).

Signature detection: In signature-based detection approach, the patterns of the abnormal action is storage in database. Signatures term is refer to the pattern of these abnormal activities. In some time, it is called misuse-based detection. The downside of this procedure is it recognizes the known assaults as it were. It favorable position of this system is it underpins the quick identification and low rate of false alarm (Egele *et al*., 2015; Shah and Singh, 2015) (Fig. 1 and Table 1).

Table 1: Comparison between behavior detection and signature detection

Techniques	Advantages	Disadvantage
Signature-based detection	Higher detection rate, Accuracy for known behaviors	Can detected only known malicious
	Simplest and effective method	Need regular update of the rules which are used
	Low false alarm rate	Rat of missing value is high
Behavior-based detection	Can test unknown and more complicated malicious	Need to be trained and tuned mod carefully, otherwise and tend to false-positives
	Detect new and unforeseen vulnerabilities	High false alarm rate and Low detection rate

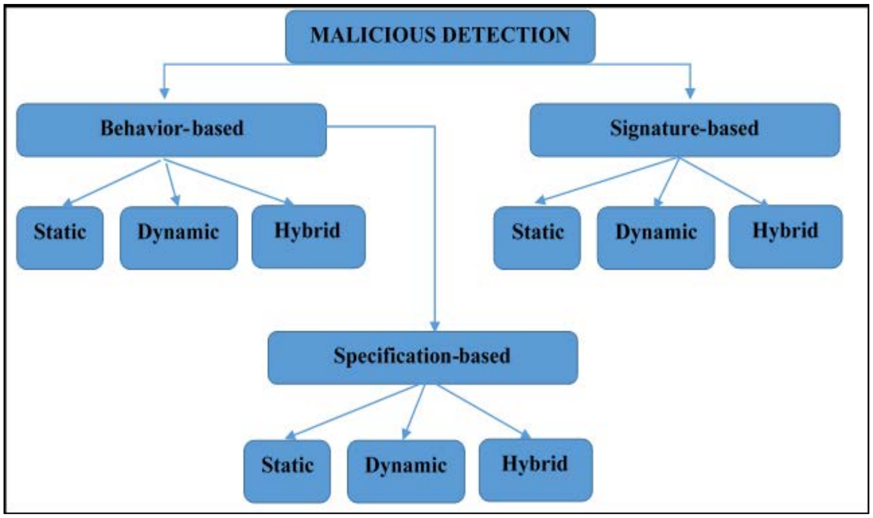

Fig. 1: Organization of malicious detection

Some types of malicious: In the following, the portion of the key classifications of malicious are briefly surveyed:

- Spyware is any innovation that guides in social event data around a man or organization without their insight
- Virus is a project or programming code that reproduces by being replicated or starting its duplicating to another system, PC boot sector or archive
- Worm is a self-recreating virus that does not change records but rather copies itself
- Logic bomb is modifying code, embedded surreptitiously or purposefully, that is intended to execute (detonate) under conditions, for example, the omission of a specific measure of time or the disappointment of a project client to react to a system charge
- Trapdoor is a technique for accessing some a part in a framework other than by the typical strategy (e.g. obtaining entrance without supplying a watchword)
- Trojan horse is a system in which malignant or destructive code is contained inside obviously innocuous programming or information in a manner that it can gain power and do its select type of harm
- RATs (Remote Admin Trojans) are an uncommon type of Trojan Horse that permits remote control over a machine
- Malware short to "malicious software" is any system or document that is unsafe to a PC client
- Mobile malicious code is web archives frequently have server-supplied code connected with them that executes inside the web browser
- Malicious font site page text that endeavors the default strategy used to de-compress Embedded in Windows (Open Type Fonts) based projects including Internet Explorer and Outlook
- Rootkits are an arrangement of programming instruments utilized by a gatecrasher to pick up and keep up access to a PC framework without the client's knowledge (Fig. 2 and 3)

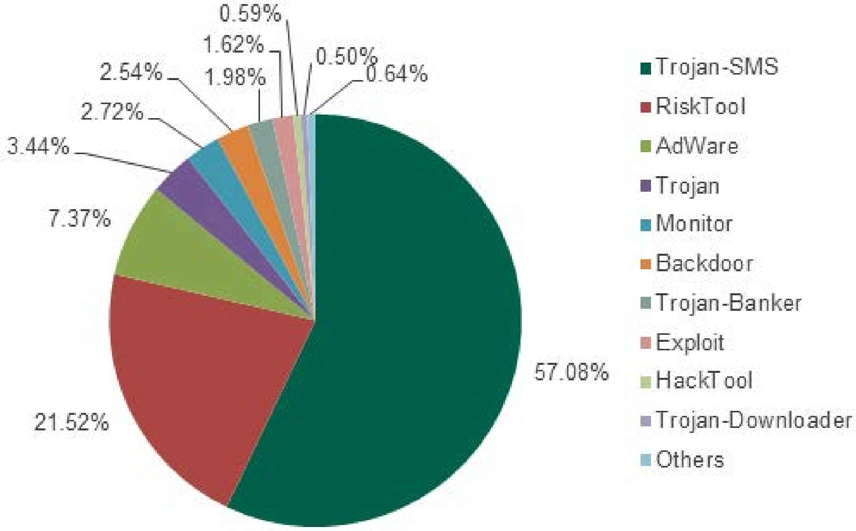

Fig. 2: Distribution of malicious types: Top 10 most active malicious 2013-2014

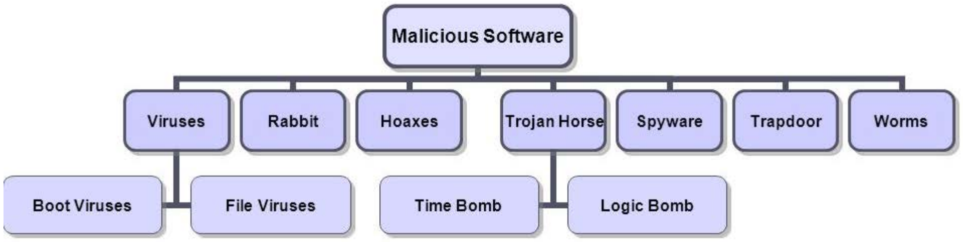

Fig. 3: Malicious categories

RESULTS AND DISCUSSION

Data mining technology based malicious detection system: Data Mining Technology has ascended as a method for recognizing patterns and trends from enormous amounts of data. It is a withdrawal of disguised prescient information or gaining from immense databases Fig. 4. Characterizes the procedure of data mining. The initial step is the aggregation of data. After that the mining operation has been executed on the data and gets the outcome. The entire system of data mining is the rehashed execution of these three stages. As the collection of data is exceptionally intricate on the grounds that the data originates from various procedures or administrations like log file, alarm massages and so on system action is enormous, so the data analysis is very hard. The data mining innovation has the capacity of removing vast databases; it is of extraordinary significance to utilize data mining techniques in malicious detection.

The important advantages of data mining methods in malicious detection system are it distinguishes the abnormal and normal data from tremendous crude data. Diverse information mining strategies are utilized to get the exact results. These data mining systems helps in get the examination amongst normal and malicious data by gaining a model. Additionally, this is not use that different data mining methods is applied on various data; rather same data mining procedure is utilized to prepare the diverse data. Various specialists to get the exact results have utilized different data mining procedures. These methods are clustering, classification, association rule mining etc (Shah and Singh, 2015).

Classification: A Classification is method of foreseeing a worth into classes i.e. class. It takes the little information and assigns it into a specific class. It separates models characterizing essential information classes. All there models are called classifiers. Classification depends on foreseeing a result based on input. A relating calculation of grouping procedures the preparation information and gets the outcomes or objectives. This objective is likewise called prediction attribute. The fundamental point of calculation is to construct the relationship between the information results and make it conceivable to foresee the outcomes.

The expectation set which contains the same arrangement of traits, the calculation takes the sources of info and produces the outcomes. The precision rate of forecast characterizes that relating calculation is great or not. The comparison of various classification strategies has been finished by the authors of Desale *et al.* (2015).

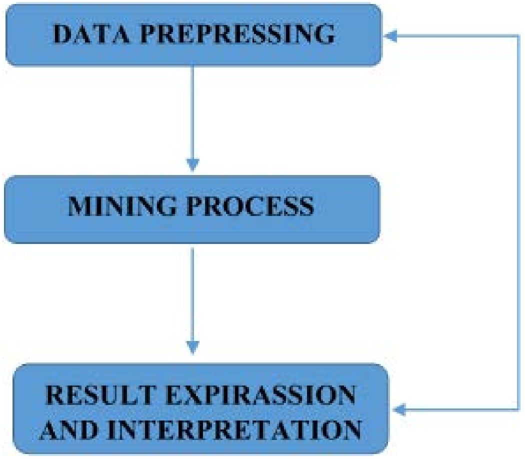

Fig. 4: Data mining processing

Classification is utilized for malicious and misuse recognition, however it is all the more ordinarily utilized for misuse discovery (Gandotra *et al*., 2014).

Clustering: A clustering is a method of making classes of the occasions of same features. These classes are known as the clusters. The features of occurrences in one bunch are not the same as the features of other group. Such as classification, technique is less proficient in the field of malicious detection. Since the measure of open system information is excessively boundless, human naming is dull and costly. From now on clustering techniques can be helpful for arranging system information for identifying malicious (Bhaya and Manaa, 2014).

Association rule: An association rule mining is a rule which suggests certain connections partner among an arrangement of items in the database. Association rule has two sections, a consequent and an antecedent. A precursor is a thing found in the information. A result is a thing that is find in a blend with the antecedent. Association rules are consequently made by analyzing the information for nonstop patterns and utilizing the criteria confidence and support to recognize the most relationships. Apriori was the main versatile calculation delivered for association rule mining.

An association rule mining finds intriguing affiliation or connection connections among a huge arrangement of information items. With Huge data constantly being gathered and saved, numerous ventures are getting to be keen on mining association rules from their databases. The relationship between datasets can be spoken to as association rules. The association rules are communicated by $I \Rightarrow S$, where S and I contain an arrangement of characteristics. This implies if a tuple satisfies I, it is likewise prone to satisfies S (Nancy *et al*., 2016).

Table 2: Comparison of various data mining techniques for malicious detection

Techniques	Algorithms	Description	Advantage	Disadvantage
Classification	Classify all the system activity into either malicious or normal. It takes each item of a dataset and allocates it to a particular class	Genetic Algorithm, Support Vector Machine (SVM), Neural Network, K Nearest Neighbor, Fuzzy Logic, Decision Tree etc.	Useful for malicious detection, analytical modeling, retailing, manufacturing. Used in both misuse and malicious detection.	Less accuracy of malicious detection
Clustering	Clustering is the way toward naming data and allotting into gatherings i.e. clustering is a division of data into gatherings of similar items	k-mean, k-mediod, PAM$^{a.1}$, CLARA$^{a.2}$, CLARANS$^{a.3}$, BIRCH$^{b.1}$, ROCK$^{b.2}$, CURE$^{b.3}$, DBSCAN$^{c.1}$, OPTICS, DENCLUE$^{c.2}$, Wave-cluster, STING (statistical information grid)	Superior to classification echnique. It can detect complex malicious over a different time.	Clustering incorporates reliance on initial centroids, reliance on number of clusters and degeneracy
Association rule	Association rules are made by separating the data for successive contexts and utilizing the criteria confidence and support to recognize the most essential relationships.	FP-Growth, Apriori etc.	Utilize when exchanging expends time. Utilize for massive datasets.	A massive number of detected rules. Getting non-intriguing rules

Comparison: As in Literature Review, analysts use distinctive techniques to detect malicious. The comparison of various techniques are explain in Table 2. Main categories of clustering data mining algorithms (Bhaya and Manaa, 2014):

- Partitional clustering algorithm
  - (PAM): Partitioning around medoids
  - (CLARA): Clustering large applications
  - (CLARANS): Clustering Large Applications based on Randomized search
- Hierarchical clustering algorithm
  - (BIRCH): Balanced Iterative Reducing and Clustering Using Hierarchies
  - (ROCK): Robust Clustering using links)
  - (CURE): (Clustering Using Representative)
- Density clustering algorithm
  - (DBSCAN): (Density-based spatial clustering of applications with noise)
  - (DENCLUE): DENsity-based clustering
- Grid-based clustering algorithm (Table 1)

Evaluation measurement of malicious detection system: There are some essential variables which are utilized to evaluate measurement of malicious detection system.

True Positive (TP): The aggregate number of typical data that are recognized as an ordinary data amid malicious detection process.

True Negative (TN): In malicious recognition, number of identified anomalous data which are really irregular data in dataset.

False Positive (FP): False alarm, all out number of recognized ordinary data yet they are real malicious.

False Negative (FN): Number of distinguished strange cases however, in genuine they are normal data.

Evaluation measurement of malicious detection system is measured in terms of detection rate, accuracy and false alarm rate:

$$\text{Detection Rate (DR)} = \left(\frac{\text{TP}}{\text{TP+FN}}\right)\times 100\%$$

$$\text{False Alarm Rate (FAR)} = \frac{\text{FP}}{\text{Number of Malicious}}$$

$$\text{Accuracy} = \left(\frac{\text{TP+TN}}{\text{TP + TN + FP + FN}}\right)\times 100\%$$

CONCLUSION

Utilization of Data Mining Technology in Malicious Detection System is an emerging trend. In this study, we show various data mining techniques utilized for malicious detection. The malicious detection system is combined with the data mining methods and algorithms malicious detect the dangers and give an immediate response to the client. Misuse detection methods are not adequate for identifying unknown malicious attack. For detecting unknown malicious attack, we have to go for malicious detection. After the review of various papers, we conclude that the most commonly utilized data mining systems are Classification, Clustering and Association Rules.

REFERENCES

Assad, A. and K. Deep, 2016. Applications of Harmony Search Algorithm in Data Mining: A Survey. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, Pant, M., K. Deep, J.C. Bansal, A. Nagar and K.N. Das (Eds.). Springer, Singapore, ISBN: 978-981-10-0450-6, pp: 863-874.

Bhaya, W. and M.E. Manaa, 2014. Review clustering mechanisms of distributed denial of service attacks. J. Comput. Sci., 10: 2037-2046.

Christodorescu, M., S. Jha, D. Maughan, D. Song and C. Wang, 2007. Malware Detection. Springer Science+Business Media, LLC., Boston, MA., Pages: 312.

Desale, K.S., C.N. Kumathekar and A.P. Chavan, 2015. Efficient intrusion detection system using stream data mining classification technique. Proceedings of the 2015 International Conference on Computing Communication Control and Automation (ICCUBEA), February 26-27, 2015, IEEE, New York, USA., ISBN:978-1-4799-6892-3, pp: 469-473.

Egele, M., T. Scholte, E. Kirda and C. Kruegel, 2012. A survey on automated dynamic malware-analysis techniques and tools. ACM Comput. Surveys, Vol. 44. 10.1145/2089125.2089126.

Gandotra, E., D. Bansal and S. Sofat, 2014. Malware analysis and classification: A survey. J. Inf. Sec., 2014: 1-9.

Hu, X., T. Chiueh and K.G. Shin, 2009. Large-scale malware indexing using function-call graphs. Proceedings of the 16th ACM conference on Computer and Communications Security, November 9-13, 2009, Chicago, IL., USA., pp: 611-620.

Nancy, D., S. Silakari and U. Chourasia, 2016. A survey over the various malware detection techniques used in cloud computing. Intl. J. Eng. Res. Technol., Vol.5,

Ravi, C. and R. Manoharan, 2012. Malware detection using windows api sequence and machine learning. Intl. J. Comput. Appl., 43: 12-16.

Shah, K. and D.K. Singh, 2015. A survey on data mining approaches for dynamic analysis of malwares. Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), October 8-10, 2015, IEEE, New York, USA., ISBN:978-1-4673-7910-6, pp: 495-499.

You, I. and K. Yim, 2010. Malware obfuscation techniques: A brief survey. Proceedings of the International Conference on Broadband, Wireless Computing, Communication and Applications, November 4-6, 2010, Fukuoka, pp: 297-300.