International Journal of Computer Applications Technology and Research
Volume 10–Issue 01, 35-41, 2021, ISSN:-2319–8656

Data Mining Approach for Cyber Security

Varsha P.Desai
Assistant Professor
Department of Computer Studies
VPIMSR, Sangli
India

Dr.K.S.Oza
Assistant Professor
Department of Computer Science
Shivaji University, Kolhapur
India

Dr.P.G.Naik
Professor
Department of Computer Studies,
CSIBER, Kolhapur
India

Abstract: Use of internet and communication technologies plays significant role in our day to day life. Data mining capability is leveraged by cybercriminals as well as security experts. Data mining applications can be used to detect future cyber-attacks by analysis, program behavior, browsing habits and so on. Number of internet users are gradually increasing so there is huge challenges of security while working in the cyber world. Malware, Denial of Service, Sniffing, Spoofing, cyber stalking these are the major cyber threats. Data mining techniques are provides intelligent approach for threat detections by monitoring abnormal system activities, behavioral and signatures patterns. This paper highlights data mining applications for threat analysis and detection with special approach for malware and denial of service attack detection with high precision and less time.

Keywords: Malware, Data Mining, Cyber-attack, Cyber Threat, Ransomware.

1 INTRODUCTION

Data mining techniques are implemented to extract hidden patterns from data. It is scientific research method for analysis, prediction and determine complex relationship between hidden patterns from large volume of data. Knowledge discovery in databases (KDD) process consist of data preprocessing, data cleaning, transformation, mining and pattern evaluation. In data mining classification of data into predefined labeled classes called as supervised leaning. Extracting similar behavioral patterns into different clusters form huge dataset called as unsupervised learning. The gaming technique of data mining where machine learning model is trained to take sequence of complex decisions in uncertain environment as per reward or punishments for specific moves called as reinforcement learning. Association, classification, clustering, regressions, decision tree, Naïve Bayes, Support vector machine, sequence mining, time series analysis are the basics techniques of data mining. Appropriate selection and implementation of data mining technique is depends on the type of data, size of data, complexity and outcome of prediction etc. Artificial intelligence based methods like neural network, fuzzy logic, genetic algorithms, deep learning are used for complex data analysis and prediction of hidden interesting patterns from complex real time database.

Data mining techniques provide systematic approach for discovering vulnerabilities, detection of threats, system loopholes, monitoring intruder's behavior and pattern. Passive attack signatures like scanning open network ports, eavesdropping, phishing, sniffing these passive attacks can be identified by using data mining algorithms. Whereas the active attack signatures like Denial of service attack, malware detection, ransomware detection is possible through data mining and artificial intelligent techniques. Machine learning technique potentially implemented for intrusion prevention system for identifying tricks and methods used by intruder as well as finding vulnerabilities, recording footprints of attack on specific network.

In supervised approach of data mining target variables can be determined according to IP address location, frequencies of web requests and time of requests. Machine learning model used to predict particular IP address is a part of which

attack signature. Implementation of liner and logistic regression, decision tree, support vector machine algorithms are used in supervised learning.

In unsupervised approach of machine learning there is no prediction of target variables while finding association between different patterns in datasets. Computer programs such as malware having similar operating behavioral pattern using clustering & association algorithm.

2 RESEARCH DESIGN

2.1 Type of the research: In the backdrop of above discussion the present research is an attempt to explore certain key aspects of cyber security. Hence the type of the research adopted in this present endeavor is descriptive research.

2.2. Objective of study:

To study data mining techniques for malware detection.

2.3 Scope of the study: The research work is focuses on study of cyber security, types of attacks, network vulnerability, cyber threats and mechanism for malware detection using data mining techniques.

3 RESULT AND DISCUSSION:

3.1 Malware detection using Data Mining:

Malicious computer program which causes abnormal behavior of computer applications though Virus, Trojan's, Warms called as malware. Using classification techniques in data mining malware can be detected and reported to the system administrator. Malware attack on system due to surfing infected websites, games or free apps download, download infected music files, installation of software application extensions, plugins or toolbar and so on. It is important to read warning messages before downloading any application, especially permissions while accessing email or personal data.

3.2 Malware Statistics: As per the research it is found that 80% damage to the system is due to malware attacks [3]. It is found that 92% malware delivered through email attachments. Mobile malware infection increase 54% from year 2018. Overall 98% malware targeted android devices. 99% malware entered through third party app downloads. Out of 10 payloads 7 are ransomware. Overall 18 million websites are infected by malware in one week. 90% financial institutions are targeted of malware from 2018. 40% ransomware victim paid the ransom. More than 50% ransomware attacks demands for bitcoin[17]

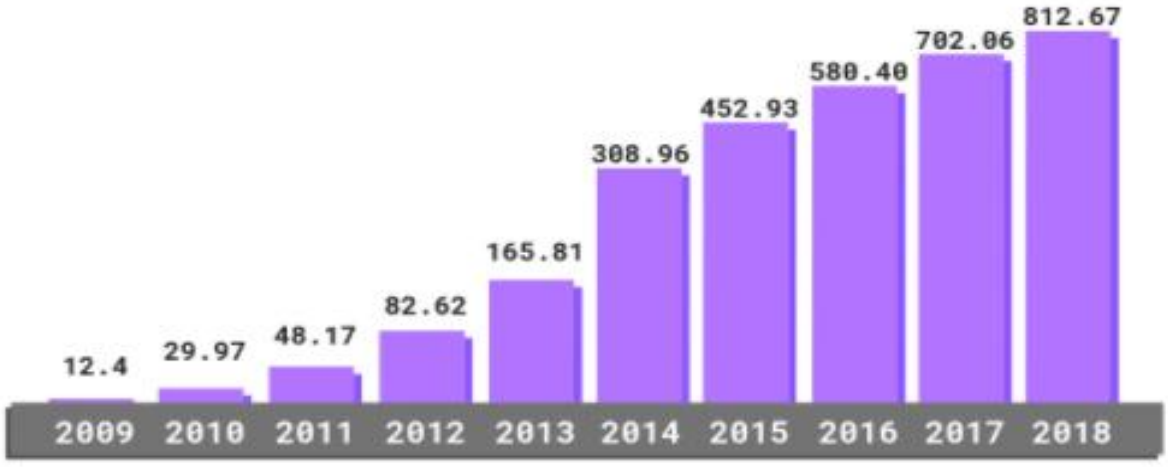

Fig.1 Total Malware Infection Growth Rate (In Millions)[17]

Now a days Malware detection is an important challenge to maintain integrity, confidentially and authentication, non-repudiation of data communicated over the internet. Data mining algorithms helps for early detection of malware as per their behavior and signature stored in database.

3.3 Malware Detection:

In behavioral based malware detection both static and dynamic analysis techniques are used for classification of program as malware. Static analysis for malware detection works on binary code which is complex to analyze and detect malware. Dynamic analysis consist of runtime code execution for testing infected files through virtual machine.[1] Malware are the malicious software code that enters into system through spam mails, email attachments, vulnerable services on internet, downloading process and browser extensions. This causes compromising computer system, unauthorized access of personal data, crippling critical infrastructure, bringing down servers, stealing system as well as network configuration information and so on. Implementation of Future extraction, classification/ clustering techniques of data mining are significant methods for malware detection [2]. Following diagram shows process of malware detection using data mining.

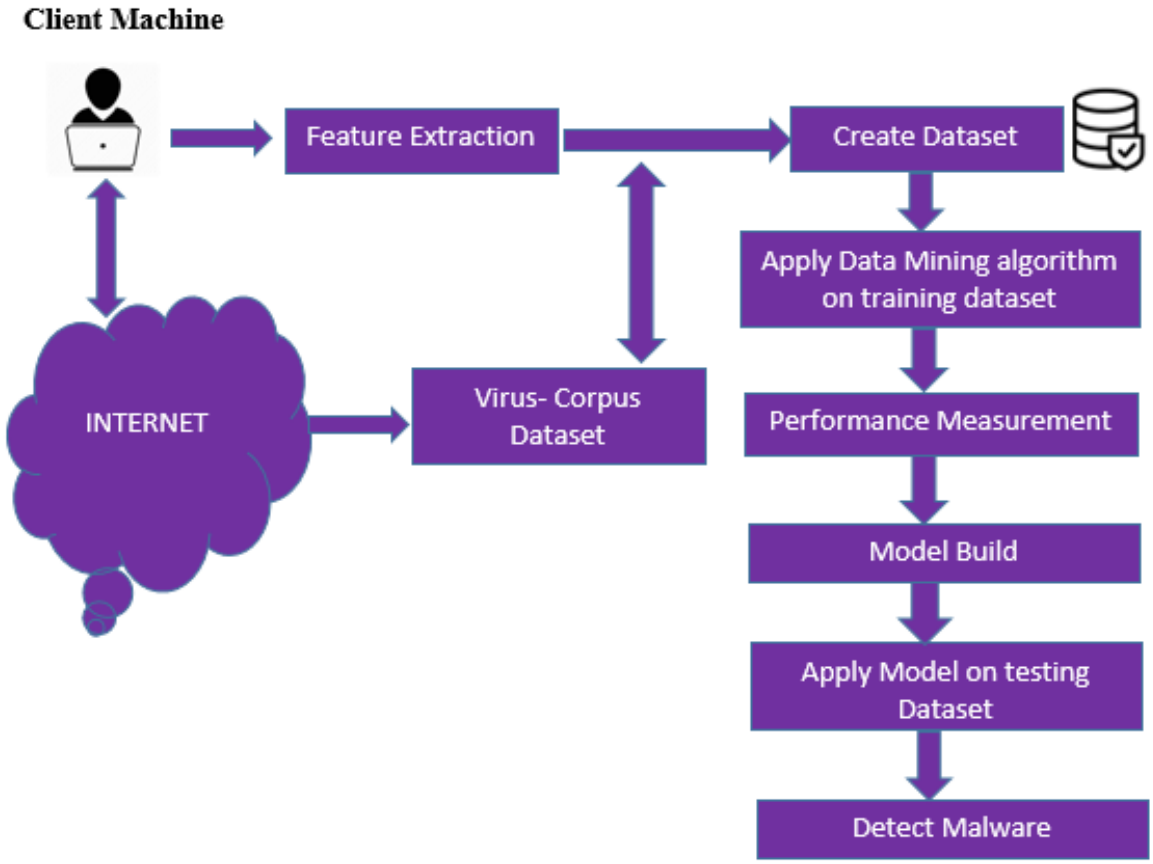

Fig.2. Malware Detection using Data Mining

When client machine is connected to internet during the scanning process machine data fetch to antimalware program. This program start future extraction process by extracting attributes from different files to create dataset. Virus files from corpus dataset is used to store virus definitions. Static analysis, dynamic analysis as well as hybrid analysis techniques are used for extracting features or patterns from data. IDA pro disassembler used to generate assembly files. Abstract assembly files are generated by eliminating operands from assembly code for better results. Extract frequent instruction association from training dataset. Data mining techniques like classification, association rule mining mechanism using Apriori algorithm are applied on training dataset based on their behavior or signature to generate frequent instructions from assembly.[5] Malware detection performance of algorithm is checked using statistical tools. Algorithm is trained until we get expected performance and finally build the model. This trained model is applied on the testing dataset to detect and report malware type and status information.

In static analysis technique of feature extraction PE files are analyses without actual execution. Detection pattern of statistical analysis in the form of windows API, N-grams, string, Opcodes or Control Flow Graph (CFG) techniques. It is one of the useful technique to investigate or explore all possible execution methods paths in malware samples [2]. Artificial neural network techniques is used to detect boot sector virus using N-gram method. Hidden dependencies between code sequences in the malware can be detected using API call method.

In Dynamic analysis debugging or profiling the code by actual execution of code at runtime. This process depend on variable value, program input, system configuration. This analysis mechanism is used for detecting new malware definitions. Detection pattern of statistical analysis in the form of debugger, simulator, emulator and virtual server based environment [2].

Hybrid analysis techniques combines benefits of static and dynamic analysis where packed malware first analyze using dynamic method and the hidden code of packed malware are extracted by comparing runtime execution of malware and its instance is analyze through static model. Hidden files are detected by dynamic analyzer while unpacked file monitor by static model [2]

3.4 Techniques of Malware Detections:

3.4.1 Signature based malware detection:

Signature database store malware footprints of previous attacks. When susceptible code is found it is tested by extracting unique bytes sequence of code as a malware signature. If it matched with existing signature the report as malware and pack malicious code file by anti-malware program. Here anti-malware program need to wait for signature until any device is victim of attack [4]. Data mining techniques like classification, regression are implemented for categorization of threat as a malware using supervised learning approach saves the time and improves the accuracy of prediction than traditional method. This method is easy to run, comprehensive malware information, search and broadly acceptable. [5] Signature database may bypass the threat using some obfuscation, cryptography methods. [4] It fails to detect the polymorphic malware that replicating information in the huge database. [5]

3.4.2 Behavior based malware detection:

Program behavior, speed of execution, response time, browsing habits, cookies information, and kinds of attachments as well as statistical properties helps to detect abnormal behavior or malicious code. In behavior based detection assembly features and API calls methods are applied using data mining algorithm. Unsupervised techniques like clustering, SVM, nearest neighboring algorithms can be implemented for behavior analysis and detection of hidden malware. This method helps to detect polymorphic malwares as well as detect data flow dependencies in the malicious software program. More time and storage space is required to detect complex behavioral pattern. Following table depicts data mining techniques for malware detection:

Type of Malware	Data Mining Techniques	Data Analysis Method
Polymorphic Malware Detection[6]	K-means	Dynamic
Android Malware Detection[7][14]	SVM, J48, Naïve Bayes	Dynamic
API Malware Detection[8]	Naïve Bays, SVM, Decision Tree, Random Forest	Dynamic
N-gram Malware Detection[9]	SVM, ANN	Dynamic
Service Oriented Mobile Malware Detection[10]	Naïve Bayes, Decision Tree	Hybrid
Sequential Pattern Malware Detection[11]	All-Nearest-Neighbor, KNN, SVM	Hybrid
Multi-objective evolutionary Malware Detection[12]	Genetic Algorithm	Static
Frequent Pattern Malware Detection[13]	Graph Mining	Static
Behavioral Malware Detection	Regression, SVM, J48	Dynamic

Table 1: Data Mining Techniques for Malware Detection

Above table depicts different data mining techniques used for malware detection according their signature and behavioral aspects. To extract hidden patterns from the data static, dynamic and hybrid data analysis techniques are used for improving accuracy of malware detection. It is the challenge for cyber security experts to select best algorithm and data analysis techniques for finding the hidden threats and provide alerts to provide data from further attacks.

4. CONCLUSION

Due to globalization usage of internet and communication technology is drastically increase. Data leakage, insecure Wi-Fi connections, lack of security awareness, hardware, software, network vulnerability are the major reasons for cybercrime. To mitigate major risk of cyber-attacks like data benches, ransomware attack, DDos attacks it is necessary to implement efficient as well as intelligent techniques for early detection of cyber threats as a proper security solution. Malware detection is one of challenge for security experts. Data mining techniques like classification, SVM, regression, decision tree, graph mining, KNN algorithms can be integrated with anti-threat system helps to detect malware before enters into system that leads to protect your IT infrastructure form further attack. Artificial neural network, genetic algorithm, deep learning mechanism provides intelligent malware detection from behavior and signature database.

REFERENCES

[1] Monire Norouzi, Alireza Souri, and Majid Samad Zamini (2016),"A Data Mining Classification Approach for Behavioral Malware Detection", Volume 2016, Journal of Computer network and Communications.

[2] Yanfang, Donald Adjeroh, et.al, (2017) "A Survey on Malware Detection Using Data Mining Techniques", ACM Computing Surveys, Vol. 50, No. 3, Article 41.

[3] Rieck. K, Willems.T, et.al (2008), Learning and classification of malware behavior, 5th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin, Heidelberg: Springer-Verlag, pp. 108–12.

[4] Sara Najari, Iman Lotfi, (2014) "Malware Detection Using Data Mining Techniques". International Journal of Intelligent Information Systems. Special Issue: Research and

Practices in Information Systems and Technologies in Developing Countries. Vol. 3, No. 6-1, pp. 33-37.

[5] Raviraj Choudhary, Ravi Saharan (2012), "Malware Detection Using Data Mining Techniques" International Journal of Information Technology and Knowledge Management, Volume 5, No. 1, pp. 85-88.

[6] Fraley JB, Figueroa M (2016) Polymorphic malware detection using topological feature extraction with data mining. In: SoutheastCon 2016, pp 1–7

[7] Sun L, Li Z, Yan Q, Srisa-an W, Pan Y (2016) SigPID: significant permission identification for android malware detection. In: 2016 11th international conference on malicious and unwanted software (MALWARE), pp 1–8

[8] Fan CI, Hsiao HW, Chou CH, Tseng YF (2015) Malware detection systems based on API log data mining. In: 2015 IEEE 39th annual computer software and applications conference, pp 255–260.

[9] Boujnouni ME, Jedra M, Zahid N (2015) New malware detection framework based on N-grams and support vector domain description. In: 2015 11th international conference on information assurance and security (IAS), pp 123–128

[10] Cui B, Jin H, Carullo G, Liu Z (2015) Service-oriented mobile malware detection system based on mining strategies. Pervasive Mob Comput 24:101–116.

[11] Fan Y, Ye Y, Chen L (2016) Malicious sequential pattern mining for automatic malware detection. Expert System Application 52:16–25.

[12] Martín A, Menéndez HD, Camacho D (2016) MOCDroid: multi-objective evolutionary classifier for Android malware detection. Soft Comput 21:7405–7415.

[13] Hellal A, Romdhane LB (2016) Minimal contrast frequent pattern mining for malware detection. Comput Secur 62:19–32.

[14] Bhattacharya A, Goswami RT (2017) DMDAM: data mining based detection of android malware. In: Mandal JK, Satapathy SC, Sanyal MK, Bhateja V (eds) Proceedings of the first international conference on intelligent computing and communication springer Singapore, Singapore, pp 187–194.

[15] Norouzi M, Souri A, Samad Zamini M (2016) A data mining classification approach for behavioral malware detection. J Comput Netw Commun 2016:9.

[16] Galal HS, Mahdy YB, Atiea MA (2016) Behavior-based features model for malware detection. J Comput Virol Hacking Tech 12:59–67. https://doi.org/10.1007/s11416-015-0244-0.

[17] Retrieved From: https://purplesec.us/resources/cyber-security-statistics/ 22 Dec 2020, 1.30pm.