International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 252-257
Published Online June 2022 in IJEAST (http://www.ijeast.com)

A SURVEY ON DETECTION METHODS USING DATA MINING

T.Baby, Dr.P.Nirmaladevi
Department of Computer Applications
Nandha Arts and Science College, Erode
Nandha Arts and Science College, Erode, Tamilnadu, India

***Abstract*- Extraction of information from huge quantities of data is known as data mining. In other words, data mining is the process of mining knowledge from data. It is a multi-disciplinary skill to extract information and evaluate future events probability. The recent applications of Data Mining are marketing, fraud detection, scientific discovery, etc. All of these will detect previously unknown although acceptable relationships among the data which are already hidden, unsuspected, and previously unknown. Nowadays, it is mainly used in information security. The significance of data mining in malware detection, intrusion detection, and fraud detection are explored in this research study.**

I. INTRODUCTION

Data mining (DM) is also referred to as data knowledge discovery (KDD). It is the process of automatically searching for large amounts of data using association rules [see Figure 1]. It is a very recent topic in computer science but uses many old computational techniques ranging from statistics, information retrieval, machine learning and method recognition. Data mining offers a few specific features to the intrusion detection scheme: [1]

- Eliminate normal functionality from alarm data to allow analysts to focus on actual attacks
- Detect faulty alarm generators and "bad" sensor signatures
- Discover the anomalous function that reveals the actual attack
- Identify long, ongoing patterns (different IP address, same function)

To perform these tasks, data miners use one or more of the following techniques:

- Data summary with statistics including locating outliers
- Visualization: Providing a graphical summary of data
- Clustering data into natural categories
- Association rule discovery: defining normal function and detecting contradictions
- Classification: Predicting which category the particular record belongs

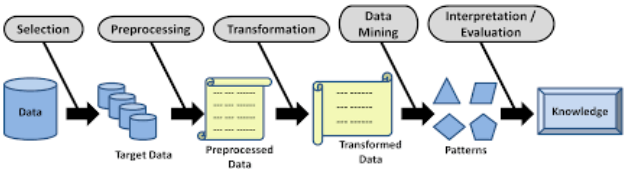


Fig. 1: Transition from raw data to vulnerable knowledge

II. DATA MINING FOR MALWARE DETECTION

Number of detection methods available today but, data mining is one of the main methods used to detect malware. The other three are scanning, activity monitoring and integrity testing. While generating a security application, developers use data mining methods to progress speed and quality of malware detection. The subsequent diagram shows procedure of malware detection using data mining.

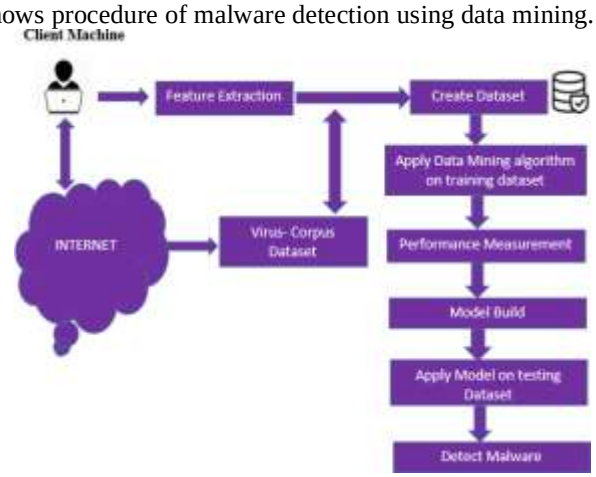


Fig.2. Malware Detection using Data Mining

There are three strategies for detecting malware:

— Anomaly Detection
— Misuse detection
— Hybrid detection

A. Anomaly detection is detection of rare events that increase suspicion by differing significantly from a large amount of data. This entails modeling default behavior of a computer or network to classify deviations from normal usage patterns. Anomaly-based detection can be used to detect prior unknown attacks

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 252-257
Published Online June 2022 in IJEAST (http://www.ijeast.com)

and to identify signatures for misuse detectors. The major problem in detection of anomaly is any deviation from the regular, still if it is a systemic behavior, will be accounted as an anomaly, hence form a high rate of false positives.

B. **Misuse detection**, also branded as signature-based detection, just identifies known attacks found on examples of signatures. It refers to detecting attacks by looking for specific patterns, such as byte series or known malicious instruction series used by malware in network traffic. This method has low rate of false positives, but does not detect latest attacks.

C. **A hybrid approach** mingles anomaly and misuse detection techniques to decrease the number of false positives at the same time raise the number of intrusion detection. It does not create any models, but rather builds a classifier using information from malicious and clean programs - a set of rules developed by a data mining algorithm. The anomaly detection system then searches for deviations from the normal profile and malware signatures on the misused detection system code.

Detection process

Malware detection consists of two steps:

- Extracting features
- Classifying/clustering

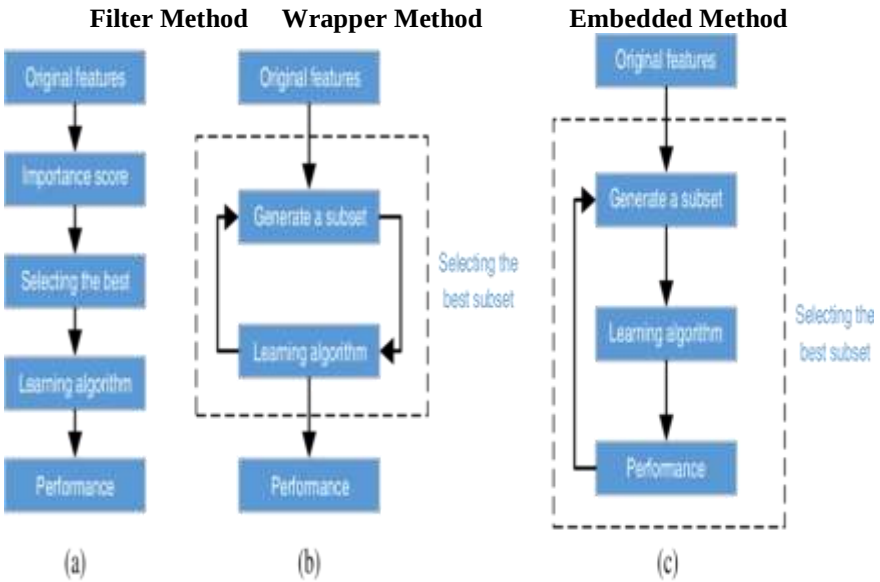
In IDS feature extraction is a significant pre-processing step. This procedure consists of feature construction and feature selection. Feature selection is the most popular method for dimension reduction. Related features are found and inappropriate ones are discarded [2]. The procedure of choosing a feature subset for further processing from the whole database was maintained in feature selection [3].
Feature selection techniques are categorized into two types, individual evaluation and subset evaluation. Feature ranking systems estimate features according to their significance and consign weights to them. Conversely, create a candidate feature that selects certain search method subset evaluation methods [4].

There are three wide groups of approaches for choosing good feature subset as filter, wrapper and hybrid approach [5].

Filter systems are commonly used for pre-processing. Selection of features is sovereign of any machine learning algorithms. Conversely, in many statistical tests, the features for their relationship with the effect variable are selected based on their scores. Filter systems do not eliminate multi colorinarity.

In wrapper methods, we try to train a model using a subset of features. Based on hypothesis made from previous model, we make a decision to insert or remove features from subgroup. These methods are usually computationally more expensive.

Embedded methods mingle the properties of both filter and wrapper methods. It is executed by algorithms with their own fixed feature selection methods.



International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 252-257
Published Online June 2022 in IJEAST (http://www.ijeast.com)

Table 1: Feature Selection Methods

Feature Selection Method	Filter Method	Wrapper Method	Embedded Method
Conception	Uses Proxy measure	Uses Predictive Model	Feature selection Is inserted in model building phase
Technique	Statistical Measures	Optimization Algorithm	Combination of Filter and wrapper method
Dataset Size	Large Dataset	Small Dataset	Small Dataset
Speed	Faster	Slower	Medium
Processing Time	Faster	Slower	Slower
Cost	Cheaper	Expensive	Expensive
Complexity	Low	High	High
Generality	High	Less	Less
Over fitting	Avoids over fitting	Prone to over fitting	Less Prone to over fitting
Performance	Sometimes may fail to select best features	Better performance	Good Performance
Example	Correlation, Chi-Square test, ANOVA, Information gain etc	Forward Selection, Backward Elimination, Stepwise Selection etc.	LASSO, Elastic Net,Ridge Regression etc.

When grouping and clustering, file samples are cataloged into groups based on feature analysis. Any classification or clustering techniques can be used to categorize samples. To organize file samples, produce a classifier using classification algorithms such as the artificial neural network (ANN), Decision Tree (DT), Support Vector Machines (SVM) or Naive Bayes (NB). Clustering is employed to group malware patterns that share related characteristics. Using machine learning techniques, every classification algorithm forms a model that signifies systematic and malicious classes.

Detect new malware by training the classifier using the file sample collection. The success of data mining techniques depends on the features to be extracted and the classification techniques used [6].

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 2, ISSN No. 2455-2143, Pages 252-257
Published Online June 2022 in IJEAST (http://www.ijeast.com)

Table 2: Data Mining Techniques for Malware Detection

Type of Malware	Data Mining Techniques	Data Analysis Method
Polymorphic Malware Detection[7]	K-means	Dynamic
Android Malware Detection[8][15]	SVM, J48, Naïve Bayes	Dynamic
API Malware Detection[9]	Naïve Bays, SVM, Decision Tree, Random Forest	Dynamic
N-gram Malware Detection[10]	SVM, ANN	Dynamic
Service Oriented Mobile Malware Detection[11]	Naïve Bayes, Decision Tree	Hybrid
Sequential Pattern Malware Detection[12]	All-Nearest-Neighbor, KNN, SVM	Hybrid
Multi-objective evolutionary Malware Detection[13]	Genetic Algorithm	Static
Frequent Pattern Malware Detection[14]	Graph Mining	Static
Behavioral Malware Detection	Regression, SVM, J48	Dynamic

III. DATA MINING FOR INTRUSION DETECTION

An IDS can supervise computer or network traffic and classify malicious activities that compromise the integrity, confidentiality, and existing information resources and prepares the system or network administrator in opposition to malicious attacks.

In addition to detecting malware code, data mining is useful to detect intrusions, examine audit results, and identify conflicting patterns. Malicious intruders occupy operating systems, networks, servers, Internet clients, and databases.

Two types of intrusion attacks:
- Host-based attacks, once intruder focuses on a particular machine or group of machines
- Network based attacks, once intruder attacks the complete network

Network-based security systems manage network flow during network firewalls, antivirus, spam filters, and network intrusion detection techniques. Host-based security systems manage incoming data on terminal through firewalls, intrusion detection techniques, and antivirus installed on host systems.

Learning is procedure of converting experience into knowledge. Based on the nature of the learning data and the relationship between the learner and the environment, learning can be divided into three categories, as noted below.

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning

Commonly used is supervised and non-supervised learning.

Supervised learning entails widening a machine learning model based on labeled data. Supervised algorithms are called supervised because the machine learning model learns from output pre-known data models. In this sense, one can consider that the entire process of learning in the supervised learning process is overseen by a supervisor. Supervised learning methods such as, Logistic Regression, Neural Networks, Support Vector Machines (SVMs), and Navie Bayes classifiers.
Supervised learning can be further classified as - regression and classification.

Regression guides and calculates continuous value response, for example forecasting real estate prices.
The classification seeks to identify the appropriate class label, such as analyzing positive / negative emotions, male

and female individuals, benign and malignant tumors, safe and unsecured loans.

The foremost idea of using unsupervised techniques is to classify malicious traffic from the normal one. Unsupervised method generally deals with cluster formation to detect intrusion. The main idea behind this is to create a set of databases and identify different behavioral datasets to identify intrusions. Unsupervised learning methods are extremely powerful tools for investigating data and classifying patterns and trends. This is the opposite of supervised learning. There is no data labeled here. They are commonly used to cluster similar input into logic groups. Some unsupervised learning algorithms include K-Means, Random Forests, hierarchical clustering.

IV. DATA MINING FOR FRAUD DETECTION

Identify different types of fraud using data mining techniques, like financial fraud, telecom fraud or computer hacking.

Credit card / financial fraud. Credit card fraud is divided into two types: offline fraud and online fraud. Offline fraud is commended by utilizing a stolen physical card in front of a store or in a call center. In most cases, the card issuer may lock the card before it is fraudulently used. Online fraud is made through internet, phone shopping or card holder- not present. Merely card details are necessary and no manual signature or card stamp is mandatory at the time of purchase.

Computer hacking. It is defined as possible chance of purposely accessing unauthorized information, manipulating information, or making a system unreliable or unusable. Intruders may be an outsider (or hacker) and an insider who knows where the computer layout, valuable data and what security provisions are in place.

Telecommunications fraud is expensive both in terms of lost earnings and exhausted capability for a network carrier. It is classified into two categories: subscription fraud and superimposed fraud. Subscription fraud takes place from receiving a subscription to a service, frequently with inaccurate identification details and with no aim of paying.

Superimposed fraud is origin by a service without the essential power identified by the look of unknown calls on a bill. This trick entails a number of ways, for example, mobile phone cloning, ghosting, internal fraud, tumbling and so on.

V. DATA MINING PROS AND CONS

Data mining has the following merits.

- Faster processing of large databases;
- Build an efficient and excusive model for all specific application case;
- Apply a few data processing techniques to spot zero day attacks.

Data mining lets you to rapidly examine huge data sets and automatically spot hidden patterns, which is essential when developing an effective anti-malware solution that detects previously unknown threats. But, the outcome of using data processing methods always depends on the superiority of the data you use.

There are some drawbacks:

- Data processing is complex, resource-intensive and expensive
- Creating an appropriate classifier can be challenging
- Malicious files must be manually inspected
- To incorporate samples of new malware, always update the classifiers
- A few data processing security issues, including the risk of reveals perceptive information without approval

VI. CONCLUSION

The internet is becoming an ever more imperative tool in everybody's life, both professional and personal, as its user and becoming more numerous. Data mining has evolved into a tool that helps its users identify vulnerabilities and provide a defensive mechanism against multiple threats to information systems.

An efficient survey prepared in the way with dissimilar data mining techniques in the circumstances of network security and intrusion detection system. When using data mining in detection process, it is essential to use only eminence data. But, preparing databases for analysis needs a lot of effort, time and resources. Clear all duplicate, incorrect and partial information before functioning with them. Lack of information or the presence of duplicate records or errors can considerably reduce the efficiency of complex data processing techniques. High quality analysis can only be confirmed if accurate and complete data are used.

VII. REFERENCES

[1] Theodoros Lappas , Konstantinos Pelechrinis "Data Mining Techniques for (Network) Intrusion Detection Systems"

[2] A. Aburomman , M. Bin Ibne Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," Appl. Soft Comput. J., vol. 38, pp. 360–372, 2016.

[3] P. Teisseyre, "Neurocomputing Feature ranking for multi-label classification using Markov networks," vol. 205,pp. 439–454, 2016.

[4] O. Y. Al-Jarrah, O. Alhussein, P. D. Yoo, S. Muhaidat, K. Taha, and K. Kim, "Data Randomization and Cluster- Based Partitioning for Botnet Intrusion Detection," IEEE Trans. Cybern., vol. 46, no. 8, pp. 1796–1806, 2016.

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

[6] 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.

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

[8] 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

[9] 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.

[10] 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

[11] 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.

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

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

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

[15] 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.