Privacy Preservation and Malware Detection Methodologies in Data Mining

Afroz Ahmed, Pranav Kala, Rachana M V N S

SRM University, Andhra Pradesh-522502

Abstract: A huge increase in cyber attacks made data mining techniques a key component to detect massive novel cyber threats. The main objective is to discover Internet threats and any other security-related threats and privacy preservation taking the guidance of domain expert and domain-specific intelligence integration in the account using the data mining methodologies.

I . Introduction

In recent times, various security incidents like malware attacks, data breaching, unauthorized access, etc have increased exponentially due to increased digitalization dependency. It is therefore important for organizations to adopt a strong relevant approach to reduce the losses caused due to cyber attacks. Data mining helps to find anomalies in a given data set. Using these patterns and correlations, you can increase revenues and reduce risks like cyber attacks, etc.

You can predict the future trends of cyber attacks by digging through the data using a broad range of techniques.

It allows you to distinguish which is the relevant data from staggering numbers of unstructured data which is 90% of the digital universe and assess that info to predicts the outcomes of cyberattacks.

Predictive analysis and data mining can give hidden insights from data which helps to track the cyber risks soon before the malware target is affected. It helps to yield automated predictions of trends and unusual behavior of the attacker.

It helps us to generate structural patterns based on the data stored which is critical for intrusion detections and attack predictions.

Since there is a large amount of data in cyberinfrastructure, many cybercriminals attempt to gain the data from the datasets through unauthorized access. And therefore, there is a need to address capabilities and challenges in cybersecurity and data mining techniques.

II . Privacy Preserving Data Mining

The goal of privacy-preserving data mining (PPDM) is to extract relevant data without effecting the sensitive data present in the database and still able to perform data mining operations efficiently.

Malicious users can use powerful techniques like data-mining and machine-learning to mine the confidential information of corporations and national departments. To preserve privacy there must be no disclosure of information or there can be a disclosure of modified information. These are the two methods for privacy concerning data mining.

Data Privacy (no disclosure of information) concentrates on the schemas of the database for the protection of sensitive data of the individuals and it reduces unauthorized access of private information, While retaining the same functions as a normal data mining technique for finding useful knowledge, and Information Privacy (disclosure of modified information) concentrates on the modification of the actual (original) information for the protection of sensitive data that can be disclosure from the database.

The main aim of the privacy-preserving data mining algorithms is to preserve personal information from exposure to the public. The privacy-preserving Data mining is all about building an algorithm that can transform original data in some manner so that both the private data and knowledge are not revealed or exposed even after a successful mining process.

II.1 Privacy-Preserving Data Mining Techniques

In terms of privacy, data mining is neutral by principle. The motive of the data mining algorithm that is used could be good or malicious which makes it useful or harmful respectively. The Privacy Preservation in Data Mining techniques are classified based on:

- Data distribution
- Data modification methods
- Rule Confusion
- Data Mining Algorithms

1.Data distribution:
Most data distribution is horizontal or vertical. We can partition a given data set in two ways horizontal and vertically. *Figure1* shows the data set that is partitioned in horizontal and vertical ways of distribution.

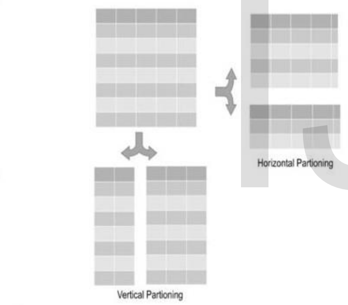

Figure1: Horizontal data distribution and vertical data distribution

In the horizontal data distribution, the data is store on different machines, and distribution is done row-wise, each site maintains complete information on a unique set of entities, The union of all the datasets is present in the integrated dataset. Respectively in Vertical data distribution, the data stored is distributed column-wise, each site maintaining different types of information and each dataset.

2.Data modification:
Data modification methods include adding noise (perturbation), Replace value with NaN(blocking), replacing several values with a statistical value (aggregation), interchanging or exchanging values (swapping), and revealing part of the available sample data (sampling) operations on the data.

3.Rule confusion:
Rule confusion refers to the balance between the data hiding and data-mining efficiency or the function using hidden data.

4.Data Mining Algorithms:
Privacy-Preserving Data mining algorithms (PPDM algorithms) deals with privacy problems caused by data mining results. The main objective of the Privacy-Preserving Data Mining is to keep private knowledge safe once the mining on the data has been done. Privacy-Preserving Data Mining methods can be analyzed from the perspective of data distribution, data modification, data mining algorithms, and rule hiding.

III . Data Mining Methodologies

Data Mining, also known as Knowledge Discovery from Data(KDD) is an automated way to extract the patterns that represent knowledge that is implicitly captured in large datasets, data warehouses, etc.

III.1 The KDD Process to extract the required knowledge and cyber activities trends and behavior from volumes of Data

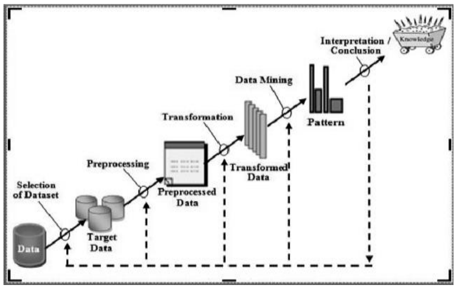

Figure2: KDD Process

In KDD, to interpret patterns from the datasets, the first step is to understand the prior knowledge and goal of the end user i.e, is to find the cyber activities trends on which the discovery is to be performed.
Figure2 gives a representation of KDD Process. A target dataset is created which is mainly focussed on a subset of variables on the dataset that is chosen on which discovery is to be performed. Data Cleaning\Pre processing is done in which the noise or inconsistent data is removed from the datasets to collect only the

necessary data for the goal to be achieved by the user.
It involves different strategies to handle the data fields that are missing.
Multiple data sources that are obtained in the pre-processing are integrated.
The data which is relevant to the analysis of the task required by the end-user is retrieved from the database that is chosen.
The retrieved data is transformed into appropriate forms using aggression techniques for data mining to be performed.
Finally, data mining operations are performed to get the required cybersecurity-related trends and data patterns.

III.2 Data Mining Techniques for analyzing the required and behavior of cyber threats.

Extracting knowledge and prediction of future trends is crucial for organizations to function without any cyber attacks. There are different data mining methods to unusual behavior and dependencies from large amounts of data.

1.Assosciation Technique

In the association technique, the pattern is identified using a transaction and a relationship among the different items in the data set in the same transaction.
It aims to observe the patterns, trends, and correlations that are frequently repeating in several databases and repositories.

2.Classification Technique

A classification technique is used to predict the patterns from an input given. It includes techniques like neural networks, statistics, and decision trees to predict unknown records. This is derived from machine learning techniques in which items in datasets are classified into pre-defined groups. It is capable of modeling datasets in such a way that items in it are classified into different classes.

3.Clustering

Meaningful object clusters are created from pieces of data which have similar characteristics. These objects are arranged in the classes that are defined by them. Clustering helps to find differences and similarities within the datasets. The data is first clustered into groups, and then similarities and differences are predicted. Based on this predicted knowledge, labels are assigned to the classes that are grouped.

4.Prediction

The prediction technique projects the data that is seen in the future. It mainly focuses on the relationship between dependent and independent variables. Usually, accurate predictions can be charted just by recognizing historical trends.

5.Sequential Patterns

It is required to detect and analyze the sequence of events or tokens that appear often in metric space. The sequential patterns technique deals with data transactions. It identifies similar patterns and trends over a while and then gives the required knowledge retrieved from the datasets.

6. Decision Trees

Decision Tree techniques automatically learn privacy-preserving and cyber-related signatures. It breaks down a complex decision-making process into simpler decision collections and provides a solution that is easier to interpret.

IV . Malware Detection Using Data Mining Techniques

Data mining is one of the main techniques used for malware detection in recent times. There are three strategies for detecting malware: anomaly detection, misuse detection, or signature-based detection.

Anomaly detection is the process of identifying unexpected items or events in data sets, which differ from the norm. An anomaly can be categorized into three types: point anomaly, contextual anomaly, and collective anomaly.
A point anomaly occurs when a tuple in a data set is far off or odd from the rest of the data.
If an anomaly occurs due to the context of observation then it is a context anomaly.
If a set of data instances help in finding an anomaly then it refers to the collective anomaly.
There are a range of features that can be used for detection: OS and application level observables such as system calls and network traffic monitoring.
Malware infection comprises two stages: exploitation and takes over stages.
Typically, in the exploitation phase, an adversary tries to hijack the control of the program execution by using a bug. Then, the malicious code is executed which takes over the machine. The main task is to detect the attack in the exploitation stage. The anomaly detection of malware arises from the observation that the malware, during execution, alters the original program flow to execute non-native code in the victim's program. Such unusual code execution

causes perturbations to the normal execution flow and flags anomaly.
In a large data network, detection of malicious or anomalous traffic becomes a complex task. It becomes difficult to monitor the huge volume of network traffic.

Misuse detection or signature-based detection is an approach in which malware is detected by past activities or pre-identified signatures. It is highly accurate with known attacks but can be toggled easily with slight modifications in the signature. Hence it is less effective than anomaly-based detection.
The signatures include patterns of log files or data packets that were found to be malicious. Log files consist of signatures that exhibit a unique pattern. The detection also arises for unauthorized access and improper file transfer protocol. The detection system executes an algorithm that attempts to match signatures with new activities to detect any attack. If the signature of the new activity matches the attack signature database, then the system raises an alert. After the detection, the system invokes certain security modules to defend against the attack. In this detection system, various machine learning techniques can be used to extract different data patterns from raw data.
However, this detection system has a drawback of raising false alarms. Hence the efficiency of this system depends on the knowledge of the known attacks.

V . Challenges of Data Mining

Using data mining techniques for cybersecurity lets large files to process faster and detect zero-day attacks.
But there are drawbacks which are to be taken into account.
Potentially malicious files are supposed to be manually inspected.
There is a risk of unauthorized disclosure of sensitive information.
Building a classifier is a challenge and these classifiers which include the new malware must be constantly updated. It is mandatory to use only quality data when data mining in cybersecurity is used.
Duplicate records and lack of information decreases the effectiveness of complex data mining techniques in cybersecurity.

VI . Conclusion

Privacy Preservation and malware detection is of course a big data processing and analytics problem. This data-driven framework is a systematic approach to address these issues. Data Mining has the potential to address such issues and lets you analyze huge datasets and extract Knowledge from it.
However, one of the technical challenges is the way unsupervised and supervised learning and the models of cyber attacks and privacy preservation are interwoven together into a robust cyberattack and privacy detection system.

VII . References:

[1] Dua, Sumeet & Du, Xian. (2011). *Data Mining and Machine Learning in Cyber Security,* published by Auerbach Publications, Boca Raton.

[2] Niranjan, A & Deepa Shenoy, P. (2016). *Security in Data Mining- A Comprehensive Survey, vol 16,* published by Global Journals Inc.(USA)

[3] Jiawei Han and Micheline Kamber (2006), *Data Mining Concepts and Techniques*, published by Morgan Kauffman, 2nd ed.

[4] J. Malone., K. McGarry, and C. Bowerman. (2006).“*Automated trend analysis of proteomics data using an intelligent data mining architecture,” in proc*. Expert Systems with Applications 30 Conf.,

[5] Jaydip Sen and Sidra Mehtab.(2020, June 19th). *Machine Learning Applications in Misuse and Anomaly Detection,* retrieved from https://www.intechopen.com/books/security-and-privacy-from-a-legal-ethical-and-technical-perspective/machine-learning-applications-in-misuse-and-anomaly-detection

[6] Yatsenko, Maria. (2018, March 29). *Using Data Mining Techniques in Cyber Security Solutions,* retrieved from https://www.apriorit.com/dev-blog/527-data-mining-cyber-security