A SURVEY OF DATA MINING TECHNIQUES FOR CYBER SECURITY

[1]Prof. K.P.Barabde, [2]Prof. V. Y. Gaud,
[1]Assistant Professor, [2] Assistant Professor
[1]Computer Science and Engineering,
[1]College of Engineering and Technology, Akola, India

Abstract:Cyber security is the area that deals with protecting from cyber terrorism. Cyber attacks include access control violations, unauthorized intrusions, and denial of service as well as insider threat. Security of an information system is its very important property, especially today, when computers are interconnected via internet. Because no system can be absolutely secure, the timely and accurate detection of intrusions is necessary. For this purpose, Intrusion Detection Systems (IDS) were designed. The IDS in combination with data mining can provide the security with next level. Data mining is the process of posing queries and extracting patterns, often previously unknown from large quantities of data using pattern matching or other reasoning techniques. This Paper gives the over view of the different data mining techniques which can be used in Cyber security for intrusion detection.

***IndexTerms* – Cyber security, Intrusion Detection System.**

I) INTRODUCTION

Cyber security is concerned with protecting computer and network systems from corruption due to malicious software including Trojan horses and viruses. Data mining for cyber securityapplications For example, anomaly detection techniques could be used to detect unusual patterns and behaviors. Data mining is the process of identifying patterns in large datasets. Data mining techniques are heavily used in scientific research as well as in business, mostly to gather statistics and valuable information to enhance customer relations and marketing strategies. Data mining has also proven a useful tool in cyber security solutions for discovering vulnerabilities and gathering indicators for baseline. In this paper, we will focus on Data mining application for cyber security. To comprehend the mechanism to be adopted in order to safeguard the computers and network, it is imperative to understand the types of threats that endanger the cyber network.

1.1 Cyber Security

Cyber security is set of rules and technologies which are mean to protect our systems, network, and data from unauthorized access, attacks, and unwanted interrupts. They are aim to maintain the confidentiality, integrity, and availability of information and information management systems through various cyber defense systems. To secure cyber infrastructure against potentially malicious threats, a growing collaborative effort between cyber security professionals and researchers from institutions, private industries, academia, and government agencies has engaged in exploiting and designing a variety of cyber defense systems.

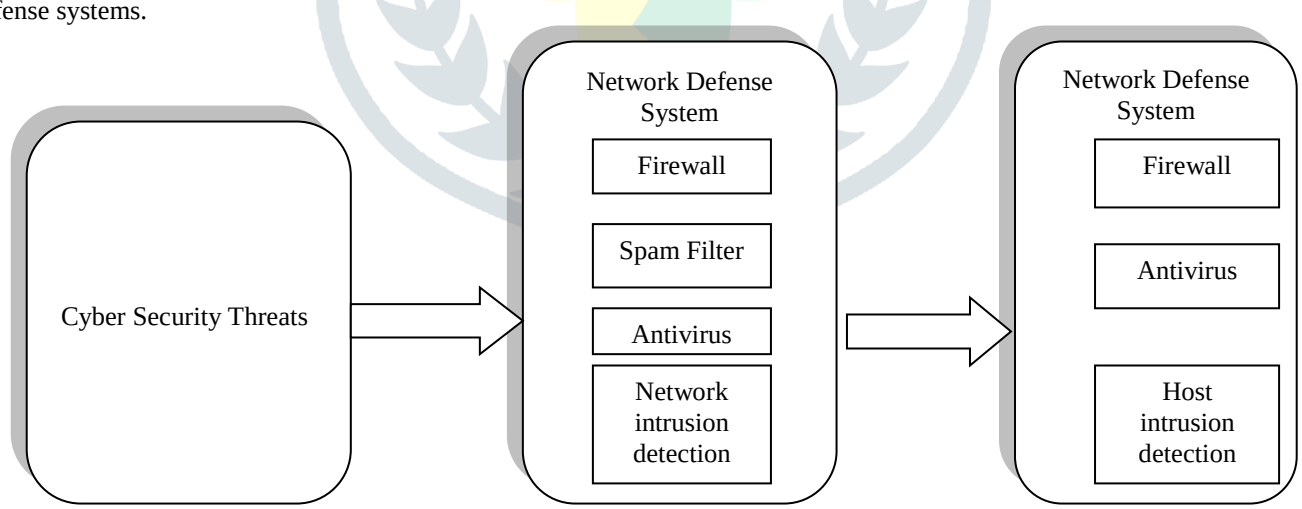

Fig. 1.1: Conventional cyber security system

Cyber security systems are composed of network security systems and host security systems. Each of these has, firewall, antivirus software, and an intrusion detection system (IDS). IDS discover, determine, and identify unauthorized use, duplication, alteration, and destruction of information systems [1].

The second line of cyber defense is composed of reactive security solutions, such as intrusion detection systems (IDSs). IDSs detect intrusions based on the information from log files and network flow, so that the extent of damage can be determined, hackers can be tracked down, and similar attacks can be prevented in the future. Data mining or knowledge discovery (KDD) is a method used to analyze data from a target source and compose that feedback into useful information. In cyber security Data mining techniques are being used to identify doubtful conditions.

Cyber Terrorism, Threats and Malicious Software

Now a days internet has allowed for a vast exchange of information. Thus has created a cyber space in which terrorists can implement attack. Cyber-terrorism, according to the O' Leary (2010) is committed through the use of cyberspace or computer resources. This use of cyber space results in there no longer being simply a physical threat of terrorism. Janczewski, & Colarik (2008) defines cyber terrorism as: "Cyber terrorism means pre-mediated, politically motivated attacks by sub national groups or clandestine agents or individuals against information and computer systems, computer programs, and data that results in violence against non-combatant targets." Cyber Terrorism is one of the major threat to world now. Over recent decades, it has become apparent that our society is becoming increasingly information technology dependant.

For Example of banking system. If terrorist attack such a system and deplete accounts of funds, then the bank could lose millions or billions of dollars. Crippling the computer system millions of hours of productivity could be lost, which is ultimately equivalent to money loss. Even a simple power failure at work could cause several hours of productivity loss which ends in financial loss. Therefore, it is imperative that our information system could be secured. Threats can occurs from outside or inside of an organization. Malicious software are the codes or procedures or programs which are mean to damage the systems, networks, clients and servers, databases. The most common types of this are virus, warms, trojan horses. Intruders try to tap into network and get vital information. It can be a human or malicious software set by humans.

1.2 Data Mining

In general, it is a process that involves analyzing information, predicting future trends, and making proactive, knowledge-based decisions based on large datasets. It is a process that involves scanning the information, predicting future trends, and making the knowledge-based decisions based on large datasets. Data mining According to Silltow (2012) automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends.

While the term data mining is usually treated as a synonym for Knowledge Discovery in Databases (KDD), it's actually just one of the steps in this process. The main goal of KDD is to obtain useful and often previously unknown information from large sets of data.

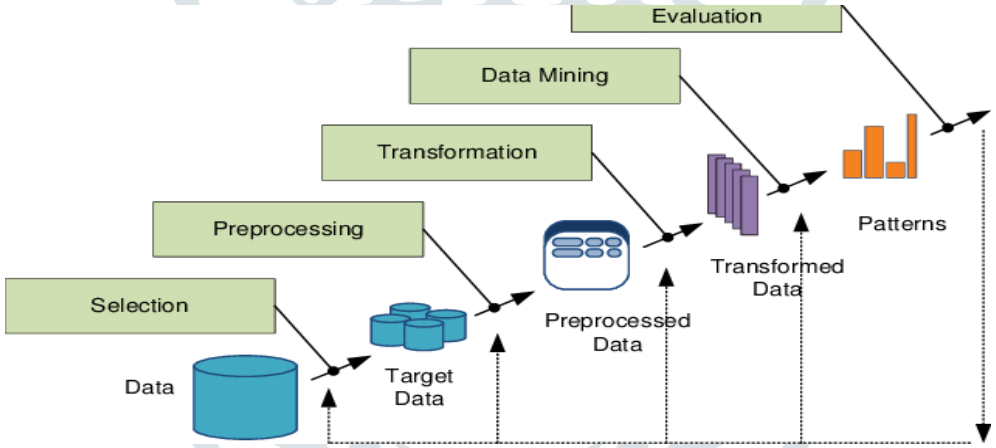

Fig 1.2: Knowledge Discovery in Database

Due to the availability of large amounts of data in cyber infrastructure and increasing number of cyber criminals attempting to gain unauthorized access to the data, there is need of capabilities to address the challenges of cyber security. Data mining tools predict future trends and behaviors by reading through database for hidden patterns, learning these behaviors is important, as they can identify and describe structural patterns which helps to generate the knowledge on the basis of that data, and helps the organization to answer the questions that were too time-consuming perversely. Learning user patterns or behaviors is critical for intrusion detection and attack predictions.

Data mining application for cyber security is the use of data mining techniques to detect cyber threats. Data mining with the combination of machine learning is being applied to problems areas such as intrusion detection and auditing in cyber security and which is very effective technique. In recent years, many IT industry giants such as Comodo, Symantec, and Microsoft have started using data mining techniques for malware detection.

DATA MINING METHODS FOR CYBER

This section describes the different DM methods for cyber security. Many methods are used for mining big data, but the following eight are the most common. Each technique is described with some detail, and references to seminal works are provided.

2.1 Association Rule

The association rule mining finds the relation among variables in database. Let's take an example IF (A AND B) THEN C. This rules describes that IF A and B are present, then there is also presence of C. Association rules have metrics that tell how often a given relationship occurs in the data.

Association Rule Mining was introduced by Agrawal et al. [2] as a way to discover interesting co-occurrences in supermarket data. It finds frequent sets of items (i.e., combinations of items that are purchased together in at least N transactions in the database), and from the frequent items sets such as {X, Y}, generates association rules of the form: $X \rightarrow Y$ and/or $Y \rightarrow X$.

A simple example of an association rule pertaining to the items that people buy together is:

IF (Bread AND Butter) → Milk

The example says that If a person buys bread and butter, then he also buy milk

The study by Brahmi [3] is a example of association rules applied on the DARPA 1998 data set to draw the relationships between TCP/IP parameters and attack types. The work based on multidimensional Association Rule Mining, in which there is more than one variables in the rules, such as (IF (service AND src_port AND dst_port AND num_conn) THEN attack_type), which is an example of a four-dimensional rule. The best results are using six dimension rules. The approach is promising for building attack signatures.

2.2 Clustering

Clustering is used to assign the similar data object in groups called clusters so that the objects in one cluster are more similar to each other than objects in other clusters. In simple word this process is used to identify data items that have similar characteristics. Clustering [4] is a set of techniques for finding patterns in high-dimensional unlabeled data. The main advantage of clustering for intrusion detection is that it can learn from audit data without requiring the system administrator to provide explicit descriptions of various attack classes.

2.3 The decision tree technique

The decision tree is a tree like structure having leaves which represent the classification and branches which represent the conjunction of features that lead to those classifications. Decision tree depends on if–then rules, but it does not requires parameters and metrics. This simple and interpretable structure allows decision trees to solve multi-type attribute problems. Decision trees can also manage missing values or noise data. However, they cannot guarantee the optimal accuracy that other machine-learning methods can.[5] The advantages of decision trees is simple implementation.

2.4 The neural network

Neural Networks are inspired by the brain and composed of interconnected artificial neurons capable of certain computations on their inputs [6]. The input data to the first layer activate the neurons of the network whose output is the input to the second layer of neurons in the network.

Neural networks re long training times and are therefore more suitable for applications where this is feasible. In Intrusion detection system Two kind of NNs are used;

- multilayered feedforward neural networks
- Kohonen's self-organizing maps.

These techniques are used to model complex relationships between inputs and outputs and to discover new patterns. The combination of Self organizing map and back propagation neural network supply a very efficient mean for detection of new intrusions.

2.5 Statistical techniques

Statistical-based systems (SBIDs) take a different approach to intrusion detection. The concept of the SBID system is simple: it determines "normal" network activity and then all traffic that falls outside the scope of normal is flagged as anomalous (not normal). It involves the collection of data relating to the behavior of legitimate user over a time period. Then statistical tests are applied to observed behavior to determine high level of confidence whether that behavior is not legitimate behavior. It fall into Two broad categories:

- Threshold detection
- Profile-based anomaly detection

This process of traffic analysis continues as long as the SBID system is active, so, assuming network traffic patterns remain constant, the longer the system is on the network, the more accurate it becomes.

DATA MINING FOR MALWARE DETECTION

Data mining is one of the four detection methods used today for detecting malware. The other three are scanning, activity monitoring, and integrity checking.

When building a security app, developers use data mining methods to improve the speed and quality of malware detection as well as to increase the number of detected zero-day attacks.

There are three strategies for detecting malware:

- Anomaly Detection
- Misuse detection
- Hybrid detection

Anomaly detection is the identification of rare events or observations which raise suspicions by differing significantly from the majority of the data. It involves modeling the normal behavior of a system or network in order to identify deviations from normal usage patterns. Anomaly based detection can also detection the previous unknown attacks and use for defining the signature for misuse detectors. The main problem with anomaly detection is that any deviation from the normal, even if it is a legitimate behavior, will be reported as an anomaly, thus producing a high rate of false positives.

Misuse detection, also known as signature-based detection, identifies only known attacks based on examples of their signatures. It refers to detection of attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware. This technique has a lower rate of false positives but can't detect zero-day attacks.

A hybrid approach combines anomaly and misuse detection techniques in order to increase the number of detected intrusions while decreasing the number of false positives. It doesn't build any models, but instead uses information from both harmful and clean programs to create a classifier – a set of rules or a detection model generated by the data mining algorithm. Then the anomaly detection system searches for deviations from the normal profile and the misuse detection system looks for malware signatures in the code.

Detection process

When using data mining, malware detection consists of two steps:

- Extracting features
- Classifying/clustering

Machine learning algorithms learn the patterns from fixed length feature vectors, and therefore feature extraction is the first step before using these algorithms for malware analysis. For features that are in the form of sequences, such as sequences of code bytes, operation codes, system calls, or any API calls, the creation of a representative feature vector is a nontrivial problem. Feature extraction can be performed by running static or dynamic analysis with or without actually running harmful software. A hybrid approach that combines static and dynamic analysis may also be used.

During classification and clustering, file samples based on feature analysis are classified into groups. To classify samples, we can use any classification or clustering techniques. To classify file samples, we need to build a classifier using classification algorithms such as Artificial Neural Network (ANN), Decision Tree (DT), Support Vector Machines (SVM) or Naive Bayes (NB). Clustering is used for grouping malware samples that shares similar characteristics. Using machine learning techniques, each classification algorithm constructs a model that represents both legitimate and malicious classes. Training a classifier using such file sample collection makes it possible to detect newly released malware. The effectiveness of data mining techniques for malware detection critically depends on the features which are extracted and the categorization techniques used.

DATA MINING FOR INTRUSION DETECTION

Apart from detecting malware code, data mining can be effectively used to detect intrusions and analyze audit results to detect anomalous patterns too. Malicious intrusions may include intrusions into operating systems, networks, servers, web clients and databases.

There are two types of intrusion attacks we can detect using data mining methods:

— Host-based attacks, when the intruder focuses on a particular machine or a group of machines

— Network-based attacks, when the intruder attacks the entire network

Network-based defense systems control network flow by network firewall, antivirus, spam filter and network intrusion detection techniques. Host-based defense systems control upcoming data in a workstation by firewall, intrusion detection techniques and antivirus installed in hosts systems.

Conventional approaches to cyber defense are mechanisms designed in authentication tools, firewalls, and network servers that monitor, track, and block viruses and other malicious attacks. For example, the Microsoft Windows operating system has a built-in Kerberos cryptography system that protects user information. Antivirus software is designed and installed in personal computers and cyber infrastructures to ensure customer information is not used maliciously. These approaches create a protective shield for cyber infrastructure. Data-capturing tools, such as Solaris BSM for SUN, Libpcap for Linux, and Winpcap for Windows, capture events from the audit files of resource information sources (e.g., network). Events can be host-based or network-based depending on where they originate. If an event originates with log files, then it is treated as a host-based event. If it originates with network traffic, then it is treated as a network-based event. A host-based event includes a sequence of commands executed by a user and a sequence of system calls launched by an application. A network-based event includes network traffic data, e.g., a sequence of TCP/IP network packets. The data-preprocessing module filters out the attacks for which good signatures have been learned.

DATA MINING FOR FRAUD DETECTION

We can detect various types of fraud using data mining techniques, it may be financial fraud, telecommunications fraud or any computer intrusions. In general, data mining techniques can be classified into two categories according to the type of the machine learning techniques for fraudulent activities it can be detected with the help of supervised and unsupervised learning.Supervised learning for fraud detection involves classification of available record in fraudulent and non-fraudulent categories. Then machines are trained to identify records according to these categories. However, these methods are only capable of identifying frauds that has already accorded [7].Unsupervised Learning for Fraud Detection method only identifies the likelihood of some records to be more fraudulent than others without statistical analysis assurance [8]. It helps in identifying privacy and security issues in data without using statistical analysis.

DATA MINING PROS AND CONS

Using data mining in cyber security lets us

- process large datasets faster;
- create a unique and effective model for each particular use case;
- apply certain data mining techniques to detect zero-day attacks.

Data mining helps us quickly analyze huge datasets and automatically discover hidden patterns, which is critical when it comes to creating an effective anti-malware solution that's able to detect previously unknown threats. However, the final result of using data mining methods always depends on the quality of data you use.

There are also certain drawbacks we need to know about:

- Data mining is complex, resource-intensive, and expensive
- Building an appropriate classifier may be a challenge
- Potentially malicious files need to be inspected manually
- Classifiers need to be constantly updated to include samples of new malware
- There are certain data mining security issues, including the risk of unauthorized disclosure of sensitive information

When using data mining in cyber security, it's crucial to use only quality data. However, preparing databases for analysis requires a lot of effort, time, and resources. You need to clean all your records of duplicate, false, and incomplete information before working with them. Lack of information or the presence of duplicate records or errors can significantly decrease the effectiveness of complex data mining techniques. Only using accurate and complete data can ensure high quality of analysis.

CONCLUSION

In this paper we have overlooked different data mining techniques for cyber security. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing.

Data mining has great potential as a malware detection tool. It allows you to analyze huge sets of information and extract new knowledge from it. When determining the effectiveness of the methods, there is not only one criterion but several that need to be taken into account. Depending on a particular IDS some might be more important than others [9]. Another crucial aspect Data mining for cyber intrusion detection is the importance of the data sets for training and testing the systems.

The main benefit of using data mining techniques for detecting malicious software is the ability to identify both known and zero-day attacks. However, since a previously unknown but legitimate activity may also be marked as potentially fraudulent, there's the possibility for a high rate of false positives.

REFERENCES

[1] A. Mukkamala, A. Sung, and A. Abraham, "Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools," in Enhancing Computer Security with Smart Technology, V. R. Vemuri,Ed. New York, NY, USA: Auerbach, 2005, pp. 125–163.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proc. Int. Conf. Manage. Data Assoc. Comput. Mach. (ACM), 1993, pp. 207–216.

[3] H. Brahmi, B. Imen, and B. Sadok, "OMC-IDS: At the cross-roads of OLAP mining and intrusion detection," in Advances in KnowledgeDiscovery and Data Mining. New York, NY, USA: Springer, 2012, pp. 13–24.

[4] K.Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[5] Sumeet Dua and Xian Du "Data Mining and Machine Learning in Cyber security"

[6] K. Hornik,M. Stinchcombe, and H.White, "Multilayer feedforward networks are universal approximators," Neural Netw., vol. 2, pp. 359–366, 1989.

[7] Bolton, R. and D. Hand, Statistical fraud detection: A review. Statistical Science 17 (3),pp. 235-255, 2002.

[8] https://www.apriorit.com/dev-blog/527-data-mining-cyber-security

[9] Anna L. Buczak, Member, IEEE, and Erhan Guven, Member, IEEE," A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016.