Review of Various Intrusion Detection Techniques based on Data mining approach

Ms.Radhika S.Landge
M.E (CSE)App
G.H.Raisoni College of Engineering
& Management,Amravati

Mr.Avinash P.Wadhe
M..E (CSE)
G.H.Raisoni College of Engineering
& Management,Amravati

Abstract

Over the past several years, the Internet environment has become more complex and untrusted. Enterprise networked systems are inevitably exposed to the increasing threats posed by hackers as well as malicious users internal to a network. IDS technology is one of the important tools used now-a-days, to counter such threats. Various IDS techniques has been proposed, which identifies and alarms for such threats or attacks. IDS are an essential component of the network to be secured. The traditional IDS are unable to manage various newly arising attacks. To deal with these new problems of networks, data mining based IDS are opening new research avenues.. Data mining provides a wide range of techniques to classify these attacks. The paper provides a study on the various data mining based intrusion detection techniques.

1. INTRODUCTION

Internet is widely spread in each corner of the world; computers all over are exposed to diverse intrusions from the World Wide Web. To protect the computers from these unauthorized attacks, effective intrusion detection systems (IDS) need to be employed. Traditional instance based learning methods for Intrusion Detection can only detect known intrusions since these methods classify instances based on what they have learned. They hardly detect the intrusions that they have not learned before. Intrusion detection techniques are of two types namely; Misuse detection and Anomaly detection. Firewalls are used for intrusion detection but they often fail in detecting attacks that take place from within the organization. To overcome this drawback of firewalls, different data mining techniques are used that handle intrusions occurring from within the organization. Data mining techniques have been successfully used for intrusion detection in different application areas like bioinformatics, stock market, web analysis etc. These methods extract previous unknown significant relationships and patterns from large databases. The extracted patterns are then used as a basis to identify new attacks. Data Mining based IDS require less expert knowledge yet provides good performance and security. These systems are capable of detecting known as well as unknown attacks from the network. Different data mining techniques like classification, clustering and association rule can be used for analyzing the network traffic and thereby detecting intrusions. [2].This paper gives a review of various data mining based techniques for intrusion detection as well as some proposed techniques and systems.

2. LITERATURE REVIEW

Intrusion detection system plays an important role in detecting malicious activities in computer systems. The following discusses the various terms related to intrusion detection. Intrusion is a type of malicious activity that tries to deny the security aspects of a computer system. It is defined as any set of actions that attempts to compromise the integrity, confidentiality or availability of any resource.

i) Data integrity: It ensures that the data being transmitted by the sender is not altered during its transmission until it reaches the intended receiver. It maintains and assures the accuracy and consistency of the data from its transmission to reception.

ii) Data confidentiality: It ensures that the data being transmitted through the network is accessible to only those receivers who are authorized to receive the respective data. It assures that the data has not been read by unauthorized users.

iii) Data availability: The network or a system resource ensures that the required data is accessible and usable by the authorized system users upon demand or whenever they need it.

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect malicious activities taking place through the network. ID is an area growing in significance as more and more sensitive data are stored and processed in networked systems. Intrusion Detection system is a combination of hardware and software that detects intrusions in the network. IDS monitor all the events taking place in the network by gathering and analyzing information from various areas within the network. It identifies possible security breaches, which include attacks from within and outside the organization and hence can detect the signs of intrusions. The main objective of IDS is to alarm the system administrator whenever any suspicious activity is detected in the network. In general, IDS makes two assumptions about the data set used as input for intrusion detection as follows: i) The amount of normal data exceeds the abnormal or attack data quantitatively. ii) The attack data differs from the normal data qualit1atively.

2.1. Major Types of Attacks

Most intrusions occur via network by using the network protocols to attack their target systems. These kinds of connections are labeled as abnormal connections and the remaining connections as normal connections. Generally, there are four categories of attacks as follows:

A. DoS – Denial of Service : Attacker tries to prevent legitimate users from accessing the service in the target machine. For example: ping-of-death, SYN flood etc.

B. Probe – Surveillance and probing : Attacker examines a network to discover well-known vulnerabilities of the target machine. These network investigations are reasonably valuable for an attacker who is planning an attack in future. For example: port-scan, ping- sweep, etc.

C. R2L – Remote to Local : Unauthorized attackers gain local access of the target machine from a remote machine and then exploit the target machines vulnerabilities. For example: guessing password etc.

D. U2R – User to Root: Target machine is already attacked, but the attacker attempts to gain access with super-user privileges. For example: buffer overflow attacks etc [2].

3. METHODOLOGY

3.1 . Techniques for Intrusion Detection

Each malicious activity or attack has a specific pattern. The patterns of only some of the attacks are known whereas the other attacks only show some deviation from the normal patterns. Therefore, the techniques used for detecting intrusions are based on whether the patterns of the attacks are known or unknown.

The two main techniques used are:

A. Anomaly Detection: It is based on the assumption that intrusions always reflect some deviations from normal patterns. The normal state of the network, traffic load, breakdown, protocol and packet size are defined by the system administrator in advance. Thus, anomaly detector compares the current state of the network to the normal behavior and looks for malicious behavior. It can detect both known and unknown attacks.

B. Misuse Detection: It is based on the knowledge of known patterns of previous attacks and system vulnerabilities. Misuse detection continuously compares current activity to known intrusion patterns to ensure that any attacker is not attempting to exploit known vulnerabilities. To accomplish this task, it is required to describe each intrusion pattern in detail. It cannot detect unknown attacks.[3]

3.2. Advantages and Disadvantages of Anomaly Detection and Misuse Detection

The main disadvantage of misuse detection approaches is that they will detect only the attacks for which they are trained to detect. Novel attacks or unknown attacks or even variants of common attacks often go undetected. The main advantage of anomaly detection approaches is the ability to detect novel attacks or unknown attacks against software systems, variants of known attacks, and deviations of normal usage of programs regardless of whether the source is a privileged internal user or an unauthorized external user. The disadvantage of the anomaly detection approach is that well-known attacks may not be detected, particularly if they fit the established profile of the user. Once detected, it is often difficult to characterize the nature of the attack for forensic purposes. Finally a high false positive rate may result for a narrowly trained detection algorithm, or conversely, a high false negative rate may result for a broadly trained anomaly detection approach.[4]

1.3 Need of Data Mining In Intrusion Detection

Data Mining refers to the process of extracting hidden, previously unknown and useful information from large databases. It is a convenient way of extracting patterns and focuses on issues relating to their feasibility, utility, efficiency and scalability. Thus data mining techniques help to detect patterns in the data set and use these patterns to detect future intrusions in similar data. The following are a few specific things that make the use of data mining important in an intrusion detection system:

i) Manage firewall rules for anomaly detection.
ii) Analyze large volumes of network data.
iii) Same data mining tool can be applied to different data sources.
iv) Performs data summarization and visualization.
v) Differentiates data that can be used for deviation analysis.
vi) Clusters the data into groups such that it possess high intra-class similarity and low inter-class similarity.

3.4. Data Mining Techniques for Intrusion Detection Systems

Data mining techniques play an important role in intrusion detection systems. Different data mining techniques like classification, clustering, association rule mining are used frequently to acquire information about intrusions by observing and analyzing the network data. The following describes the different data mining techniques:

A. Classification**:** It is a supervised learning technique. A classification based IDS will classify all the network traffic into either normal or malicious. Classification technique is mostly used for anomaly detection. The classification process is as follows: i) It accepts collection of items as input. ii) Maps the items into predefined groups or classes defined by some attributes. iii) After mapping, it outputs a classifier that can accurately predict the class to which a new item belongs.

B. Association Rule**:** This technique searches a frequently occurring item set from a large dataset. Association rule mining determines association rules and/or correlation relationships among large set of data items. The mining process of association rule can be divided into two steps as follows: i) Frequent Item set Generation Generates all set of items whose support is greater than the specified threshold called as minsupport. ii) Association Rule Generation From the previously generated frequent item sets, it generates the association rules in the form of ―if then‖ statements that have confidence greater than the specified threshold called as minconfidence. The basic steps for incorporating association rule for intrusion detection are as follows: i) The network data is arranged into a database table where each row represents an audit record and each column is a field of the audit records. ii) The intrusions and user activities shows frequent correlations among the network data. Consistent behaviors in the network data can be captured in association rules. iii) Rules based on network data can continuously merge the rules from a new run to aggregate rule set of all previous runs. iv) Thus with the association rule, we get the capability to capture behavior for correctly detecting intrusions and hence lowering the false alarm rate.

C. Clustering**:** It is an unsupervised machine learning mechanism for discovering patterns in unlabeled data. It is used to label data and assign it into clusters where each cluster consists of members that are quite similar. Members from different clusters are different from each other. Hence clustering methods can be useful for classifying network data for detecting intrusions. Clustering can be applied on both Anomaly detection and Misuse detection. The basic steps involved in identifying intrusion are follows : i) Find the largest cluster, which consists of maximum number of instances and label it as normal. ii) Sort the remaining clusters in an ascending order of their distances to the largest cluster. iii) Select the first K1 clusters so that the number of data instances in these clusters sum up to ¼`N and label them as normal, where ` is the percentage of normal instances. iv) Label all other clusters as malicious. v)After clustering, heuristics are used to automatically label each cluster as either normal or malicious. The self-labeled clusters are then used to detect attacks in a separate test dataset. From the three data mining techniques discussed above clustering is widely used for intrusion detection because of the following advantages over the other techniques:

i) Does not require the use of a labeled data set for training. ii) No manual classification of training data needs to be done. iii) Need not have to be aware of new types of intrusions in order for the system to be able to detect them.

3.5. Where to do Intrusion Detection

According to the monitored system, the source of input information can be on a host or network or host and network. Thus IDS is further classified into three categories as follows :

i) Network-based intrusion detection system (NIDS)
It is an independent platform that identifies intrusions by examining network traffic and monitors multiple hosts. Network intrusion detection systems gain access to network traffic by connecting to a network hub, network switch configured for port mirroring, or network tap.

ii) Host-based intrusion detection system (HIDS)
It consists of an agent on a host that identifies intrusions by analyzing system calls, application logs, file-system modifications (binaries, password files, capability databases, Access control lists, etc.) and other host activities and state. In a HIDS, sensors usually consist of a software agent.

iii) Hybrid Intrusion detection system (Hybrid IDS)
It complements HIDS system by the ability of monitoring the network traffic for a specific host; it is different from the NIDS that monitors all network traffic . In computer security, a Network Intrusion Detection System (NIDS) is an intrusion detection system that attempts to discover unauthorized access to a computer network by analysing traffic on the network for signs of malicious activity[3].

3.6. New techniques introduced for IDS based on data mining

3.6.1 Multi Agent Based Approach For Network Intrusion Detection

In a multi agent based approach is used for network intrusion detection. An adaptive NIDS will be used. Here more numbers of agents are used which will be continuously monitoring the data to check for any intruder which might have entered in the system. Each agent is trained accordingly so that it can check for any type of intruder entering into the system. There are five types of agent based on three data mining techniques, which are clustering, association rules and sequential association rules approaches. The problem is that current NIDS are tuned specifically to detect known service level network attacks. Attempts to expand beyond this limited realm typically results in an unacceptable level of false positives. At the same time, enough data exists or could be collected to allow network administrators to detect these policy violations. Unfortunately, the data is so volumous, and the analysis process so time consuming, that the administrators don't have the resources to go through it all and find the relevant knowledge, save for the most exceptional situations, such as after the organization has taken a large loss and the analysis is done as part of a legal investigation. In other words, network administrators don't have the resources to proactively analyze the data for policy violations, especially in the presence of a high number of false positives that cause them to waste their limited resources. An adaptive NIDS based on data mining techniques is proposed. However, unlike most of the current researches, which only one engine is used for detection of various attacks; the system is constructed by a multi-agent, which are totally different in both training and detection processes. After training with normal traffic for a network behavior, when new type of attack comes, the system can detect such anomaly by distinguishing it from normal traffic [5].

3.6.2 Intrusion detection using fuzzy logic and data mining

The method extracts fuzzy classification rules from numerical data, applying a heuristic learning procedure. The learning procedure initially classifies the input space into non-overlapping activation rectangles corresponding to different output intervals.There is no overlapping and inhibition areas. However, the disadvantage listed is, the high false positive rates which is the primary scaling of all the IDS. Researcher describes the approaches to address three types of issues: accuracy, efficiency, and usability.First issues of improving accuracy is achieved by using data mining programs to analyze audit data and extract features that can distinguish normal activities from intrusions. Second issue, efficiency is improved by analyzing the computational costs of features and a multiple-model cost-based approach is used to produce detection models with low cost and high accuracy. Third issue, improved usability, is solved by using adaptive learning algorithms to facilitate model construction and incremental updates; unsupervised anomaly detection algorithms are used to reduce the reliance on labelled data. Researchers developed the Fuzzy Intrusion Recognition Engine (FIRE) using fuzzy sets and fuzzy rules. FIRE uses simple data mining techniques to process the network input data and generate fuzzy sets for every observed feature. The fuzzy sets are then used to define fuzzy rules to detect individual attacks. FIRE does not establish any sort of model representing the current state of the system, but instead relies on attack specific rules for detection. Instead, FIRE creates and applies fuzzy logic rules to the audit data to classify it as normal or anomalous. Dickerson et al. found that the approach is particularly effective against port scans and probes. The primary disadvantage to this approach is the labour intensive rule generation process. The research work shown by Figure 3.5 can be considered as an extension of the above work by automating the rule generation process.

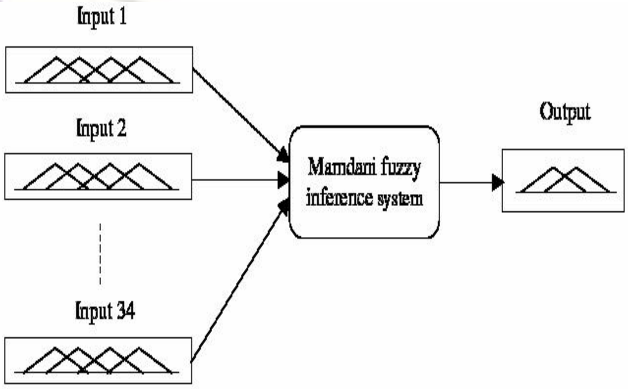

Figure 3.5 A ID model using neural networks and fuzzy logic

The model combines neural networks and fuzzy logic. This system works by mapping a template graph and user action graph to determine patterns of misuse. The output of this mapping process will be used by the central strategic engine to determine whether an intrusion has taken place or not. The major drawback is that new type attacks rules need to be given by the external security officer i.e. it does not automate rule generation process and more number of components prevents it from working fast. [6].

3.6.3 Data Mining And Real Time IDSs

Even though offline processing has a number of significant advantages, data mining techniques can also be used to enhance IDSs in real time. Lee were one of the first to address important and challenging issues of accuracy, efficiency, and usability of real-time IDSs. They implemented feature extraction and construction algorithms for labeled audit data. Eg. entropy, conditional entropy, relative entropy, information gain, and information cost to capture intrinsic characteristics of normal data and use such measures to guide the process of building and evaluating anomaly detection models. A serious limitation of their approaches (as well as with most existing IDSs) is that they only do intrusion detection at the network or system level. However, with the rapid growth of e-Commerce and e-Government applications, there is an urgent need to do intrusion detection at the application-level.

This is because many attacks may focus on applications that have no effect on the underlying network or system activities.[7]

4. SURVEY OF APPLIED TECHNIQUES

In this section we present a survey of data mining techniques that have been applied to IDSs by various research groups.

4.1. Machine Learning

Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. In contrast to statistical techniques, machine learning techniques are well suited to Clustering and Classification are probably the two most popular machine learning problems. Techniques that address both of these problems have been applied to IDSs.

4.1.1 Cla1ssification Techniques

In a classification task in machine learning, the task is to take each instance of a dataset and assign it to a particular class. A classification based IDS attempts to classify all traffic as either normal or malicious. The challenge in this is to minimize the number of false positives (classification of normal traffic as malicious) and false negatives (classification of malicious traffic as normal). Five general categories of techniques have been tried to perform classification for intrusion detection purposes:

a) Neural Networks : The application of neural networks for IDSs has been investigated by a number of researchers. Neural networks provide a solution to the problem of modeling the users' behavior in anomaly detection because they do not require any explicit user model. Neural networks for intrusion detection were first introduced as an alternative to statistical techniques in the IDES intrusion detection expert system to model . The researcher McHugh have pointed out that advanced research issues on IDSs should involve the use of pattern recognition and learning by example approaches for one reason:

- The capability of learning by example allows the system to detect new types of intrusion.

A different approach to anomaly detection based on neural networks is proposed by Lee et al. While previous works have addressed the anomaly detection problem by analyzing the audit records produced by the operating system, in this approach, anomalies are detected by looking at the usage of network protocols.

b) Fuzzy Logic : Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic. An enhancement of the fuzzy data mining approach has also been applied by Florez et al. The authors use fuzzy data mining techniques to extract patterns that represent normal behavior for intrusion detection. Luo also attempted classification of the data using Fuzzy logic rules.

c) Genetic Algorithm : Genetic algorithms were originally introduced in the field of computational biology. Since then, they have been applied in various fields with promising results. Fairly recently, researchers have tried to integrate these algorithms with IDSs.

d) Support Vector Machine : Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. SVMs attempt to separate data into multiple classes. Mukkamala, Sung, et al. used a more conventional SVM approach. They used five SVMs, one to identify normal traffic, and one to identify each of the four types of malicious activity in the KDD Cup dataset. Eskin et al. and Honig et al. used an SVM in addition to their clustering methods for unsupervised learning. The achieved performance was comparable to or better than both of their clustering methods.

4.1.2 Clustering Techniques

Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Machine learning typically regards data clustering as a form of unsupervised learning. Clustering is useful in intrusion detection as malicious activity should cluster together, separating itself from non-malicious activity. Clustering provides some significant advantages over the classification techniques already discussed, in that it does not require the use of a labeled data set for training[7].

4.1.2 Existing Systems

In this section, we present some of the implemented systems that apply data mining techniques in the field of Intrusion Detection.

a) The MINDS System : The Minnesota Intrusion Detection System (MINDS), uses data mining techniques to automatically detect attacks against computer networks and systems. While the long-term objective of MINDS is to address all aspects of intrusion detection, the system currently focuses on two specific issues:

b) EMERALD (SRI) : EMERALD is a software-based solution that utilizes lightweight sensors distributed over a network or series of networks for real-time detection of anomalous or suspicious activity. EMERALD sensors monitor activity both on host servers and network traffic streams. By using highly distributed surveillance and response monitors, EMERALD provides a wide range of information security coverage, real-time monitoring and response, protection of informational assets.

c) IDSs in the Open Market: Various systems that employ data mining techniques have already been released as parts of commercial security package.[7]

5.CONCLUSION

The application of Data Mining in Intrusion Detection System is emerging trend in the recent years. The Data Mining techniques can extract characteristics of sample data, thus reduces the difficulties involved in the collection of training data. Thereby achieving the active defence for Intrusion Detection System. The traditional Intrusion Detection System cannot do all of these. It is necessary to describe this indeterminacy because the data of network traffic and host audit and the detective process of Intrusion Detection System are indeterminable. This paper describes the distinction of attack degree due to above reason. The Data Mining plays a major role in wide variety of its application areas. The sequence representation of data in network traffic is uncertain. There is limitation in the application of intrusion detection technology. The flexibility of system is not good to analyze the huge amount of data based upon proposed method. Still there is scope for research in this area.[8]

REFERENCES

[1] Mrs. Sneha Kumari, Dr. Maneesh Shrivastava “A Study Paper on IDS Attack Classification Using Various Data Mining Techniques” International Journal of Advanced Computer Research Volume-2 Number-3 Issue-5 September-2012. [2] Mitchell D’silva, Deepali Vora “Comparative Study of Data Mining Techniques to Enhance Intrusion Detection ” International Journal of Engineering Research and Applications (IJERA) Vol. 3, Issue 1, January -February 2013.

[3] S.A.Joshi, Varsha S.Pimprale “Network Intrusion Detection System (NIDS) based on Data Mining” International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013.

[4] Reema Patel, Amit Thakkar, Amit Ganatra “A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems’ International Journal of Soft Computing and Engineering (IJSCE) Volume-2, Issue-1, March 2012

[5] Ankita Agarwal “ Multi Agent Based Approach For Network Intrusion Detection Using Data Mining Concept” Journal of Global Research in Computer Science, 3 (3), March 2012.

[6] Miss. Prajkta P. Chapke & Prof. A.B. Raut “ Intrusion Detection System using Fuzzy logic and Data Mining Technique” International Journal of Advanced Research in Computer Science and Software Engineering 2 (12), December – 2012.

[7] Monali Shetty, Prof. N.M.Shekokar “Data Mining Techniques for Real Time Intrusion Detection Systems” International Journal of Scientific & Engineering Research Volume 3, Issue 4, April-2012.

[8] Alok Ranjan, Dr. Ravindra S. Hegadi, Prasanna Kumara “Emerging Trends in Data Mining for Intrusion Detection” International Journal of Advanced Research in Computer Science Volume 3, No. 2, March-April 2012.

Ms.Radhika S. Landge has received her B.E in computer Science & Engineering from Sant Gadgebaba Amravati University (SGBAU) and persuing M.E (CSE) From G.H.Raisoni College of Engineering & Management,Amravati. Her research interest includes Network Security and Data mining.

Prof. Avinash P. Wadhe: Received the B.E from SGBAU Amravati university and M-Tech (CSE) From G.H Raisoni College of Engineering, Nagpur (an Autonomous Institute). He is Currently an Assistant Professor with the G.H Raisoni College of Engineering and Management ,Amravati SGBAU Amravati university.His research interest include Network Security , Data mining and Fuzzy system .He has contributed to more than 20 research papers. He had awarded with young investigator award in international conference.