A Survey of Data Mining Technique to implement the Anomaly Intrusion Detection System

Susheel Kumar Tiwari[1], Dr. Manish Shrivastava[2]

[1] *PhD Research Scholar Mewar University, Chittorgarh, Rajasthan, (INDIA)*

[2]*Professor & Head (CSE) L.N.C.T Bhopal, Affiliated to R.G.P.V Bhopal , M.P.,(INDIA)*

ABSTRACT

Security is one of the most challenging areas for computers and networks. Intrusion Detection System tools aim to detect computer attacks, computer misuse and to alert the proper individuals upon detection. But still they face challenges in robust and changing environment. In Information Security, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. Intrusion detection does not, in general, include prevention of intrusions. In this paper, we are mostly focused on data mining techniques that are being used for such purposes. We debate on the advantages and disadvantages of these techniques. Finally we present a new idea on how data mining can aid IDSs.

Key words: Intrusion detection system, Data Ming, Attacks

I. INTRODUCTION

Intrusions are the violations or impending threads of violations of computer security policies. Attack is any attempt to destroy, expose, alter, disable, steal or gain the unauthorized access to or make the unauthorized use of assets. Attack can be active or passive. An active attack attempts to alter system resources or affect their operations, hence comprises the integrity or availability. A passive attack attempts to learn or make use of information from the system but does not affect system resources, hence comprises confidentiality. An attempt of attack can takes place from inside or outside the organization. Insider attacker is one who has authorized access to system resources but use them in illegitimate way. Outside attacker is the illegal user of the system. A close-in attack involves someone attempting to get physically close to network components, data, and systems in order to learn more about a network. In phishing attack the hacker creates a fake web site that looks exactly like an original website, when the user attempts to log on with their account information, the hacker records the username and password and then tries that information on the real website. In a hijack attack, a hacker takes over a session between you and another person and disconnects the other person from the communication and you still consider that you are talking to the original party and may send private information to the hacker by an accident. In a spoofing attack, the hacker modifies the source address of the packets he or she is sending so that they appear to be coming from someone else. This may be an attempt to bypass firewall rules. In Exploit type of attack, the attacker knows of a security problem within an operating system or a piece of software and leverages that knowledge by exploiting the vulnerability. In Password attack, an attacker tries to break the passwords stored in a network account database or in a password-protected file. Intrusion Detection System (IDS) is the most authoritative system that can handle the intrusions of the computer environment by alerting the analyst so that they can take corrective actions to prevent that intrusion. Major functions of Intrusion detection system involves,

1. Used to examine the network traffic.
2. Identifying possible events by monitoring both user and system.
3. Logging information about user and system
4. Analyzing system configuration and vulnerabilities.
5. Assessing file and system integrity [1].
6. Recognizing abnormal activities and patterns of typical types of attacks.
7. Reporting them to network security administrator. Additional to this, many organizations use Intrusion detection system for other purposes such as identifying problems with security policies, documenting the existing threats, deterring the individuals from violating the security policies.

II. WHY WE NEED IDS?

Of the security incidents that occur on a network, the vast majority come from inside the network. These attacks may consist of otherwise authorized users who are disgruntled employees. The remainder comes from the outside, in the form of denial of service attacks or attempts to penetrate a network infrastructure .IDS tools allow for complete supervision of networks, regardless of the action being taken, such that information will always exist to determine the nature of the security incident and its sources. The main function of IDS[2] includes:

- Monitoring and analyzing the information gathered from both user and system activities.
- Analyzing configurations of system and evaluating the file integrity and system integrity.
- For static records, it finds out the abnormal pattern.
- To recognize abnormal pattern, it use static records and alert to system administrator

III. IDS TAXONOMY

There are many approaches to solutions. They are:

1. Signature based
2. Anomaly based
3. Host based
4. Stack based
5. Network based

1. Signature Based: This possess an attack description that can be matched to sense activities. Most signature based analysis system is simple pattern matching system. In this system there are many advantages and disadvantages. Some of the drawbacks in this system is:

i.They cannot detect the novel attacks.

ii.Suffer due to false alarms

iii.Have to be programmed for every new attack.

Advantages:

i.Simple to implement

ii.light weight

iii.low false positive rate

2. Anomaly Based: It observes the normal use of network as noise characterization which is distinct from noise is assumed as intrusion. There are also advantages and disadvantages in this system. Some disadvantages are: Intrusions are accompanied by manifestations that are sufficiently unusual and raises the false alarm and compromise the effectiveness of the intrusion detection system.

3. Host based: Host operating system also known as application logs in audit information. This includes event like identification and authentication, file opens and program and then analyzed to detect trails.

(a) Misuse/Signature detection: This technique looks for patterns and signatures of already known attacks in the network traffic. A constantly updated database is usually used to store the signatures of known attacks. The way this technique deals with intrusion detection resembles the way that anti-virus software operates.

4. Stack Based: These are integrated with TCP / IP stack which allows packets to be watched and traversed. TCP/IP allows the IDS to pull packets from the stack before the OS or the application has the chance to preprocess the data in packets.

5. Network Based: Network based system looks for the attacks in the analysis data and signatures checks in network traffic. A filter[3] is used to identify which system is used to discard or paused. It helps to filter out un-malicious activities.

IV.FUNCTIONS OF INTRUSION DETECTION

1. It Monitors and analyzes both user and system activities
2. Analyzes the system configurations and vulnerabilities
3. Assesses the system and file integrity
4. Has the Ability to recognize the patterns typical of attacks
5. It Analysis the abnormal activity patterns Tracks the user policy violations

V. DRAWBACKS OF IDSS

Intrusion Detection Systems (IDS) have become a standard component in security infrastructures[4] as they allow network administrators to detect policy violations. These policy violations range from external attackers trying to gain unauthorized access to insiders abusing their access. Current IDS have a number of significant drawbacks:

- Current IDS are usually tuned to detect known service level network attacks. This leaves them vulnerable to original and novel malicious attacks.

- Data overload: Another aspect which does not relate directly to misuse detection but is extremely important is how much data an analyst can efficiently analyze. That amount of data he needs to look at seems to be growing rapidly. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of records per day.

• False positives: A common complaint is the amount of false positives an IDS will generate. A false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly.

• False negatives: This is the case where an IDS does not generate an alert when an intrusion is actually taking place. (Classification of malicious traffic as normal) Data mining can help improve intrusion detection by addressing each and every one of the above mentioned problems

VI .DATA MINING. WHAT IS IT?

Data mining (DM)[5], also called Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using association rules. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition.

Here are a few specific things that data mining might contribute to an intrusion detection project:

- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify false alarm generators and ”bad” sensor signatures
- Find anomalous activity that uncovers a real attack
- Identify long, ongoing patterns (different IP address, same activity) To accomplish these tasks, data miners employ one or more of the following techniques:
- Data summarization with statistics, including finding outliers
- Visualization: presenting a graphical summary of the data • Clustering of the data into natural categories
- Association rule discovery: defining normal activity and enabling the discovery of anomalies
- Classification: predicting the category to which a particular record belongs

Knowledge is the information which can be converted into knowledge about historical patterns and future trends. The Knowledge Discovery in Database (KDD) process is generally defined with the stages

1. Selection
2. Pre-processing
3. Transformation
4. Data Mining
5. Interpretation/Evaluation[6]

Data mining is a process to extract information and knowledge from a large number of incomplete, noisy, fuzzy and random data. It is a suitable way of extracting patterns, which represents mining completely stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability.

Data mining consists of five major elements

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.

5. Present the data in a useful format, such as a graph or table.

6.1 Advantages of Data Mining Techniques

i. Problems with large databases may contain valuable implicit regularities that can be discovered automatically[7].

ii. Difficult-to-program applications, which are too difficult for traditional manual programming.

iii. Software applications that modify to the individual users preferences, such as modified advertising.

VII. A REVIEW OF LITERATURE

This section discusses about various detection algorithms for network security.

[1] Network Intrusion Detection System based on Data Mining – S.A. Joshi, et. al.[1]

In this paper the author discuss about the data mining algorithms and Intrusion detection system to detect the unknown attacks from the dataset. There different kinds of attacks but the authors of this paper discuss the few kinds of attacks. They compares the four types of attacks are: a) Probing attack b) Denial of service c) User to root d) Remote to local Then the author listed out the various data mining techniques and intrusion detection techniques which is used for the detecting the attacks like signature based detection, anomaly based detection, network- based intrusion detection system, host-based detection system.

[2] Anomaly Detection in Network using Data mining Techniques – Sushil Kumar Chaturvedi, et. al.[2]

The main work of this compares the two types of algorithms C4.5 and Support Vector Machine (SVM). First the given dataset is pre-processing and then the data can be partition into training and testing. The third stage the dataset is applied in C4.5 and SVM algorithm. The author of this paper compares these two algorithms and find out the detection rate comparison and false alarm rate comparison. By using these two data mining techniques they justify the C4.5 algorithm is better than the SVM.

[3] Application of Genetic Algorithm in Intrusion Detection System – Omprakash Chandrakar, et. al.[3]

This paper describes about basic concepts of network intrusion detection system, components and types of attacks. The IDS contains the three types of components namely data source, analysis engine, response manager. This paper gives the overview of genetic algorithm. The genetic algorithm randomly selected the input (chromosome) and calculates the fitness value for each generated initial chromosome. The iteration has performed some specific operations namely sorting, selection, crossover, mutation and finally calculates the fitness value for chromosome.

[4**] Anomaly Detection System by Mining Frequent Pattern using Data Mining Algorithm from Network Flow – A.R. Jakhale, et. al**.[4]

This paper describes an anomaly detection system and its two phases namely training and testing. The sliding window and clustering is used to monitoring the network traffic by mining the frequent patterns using algorithms. The algorithms are so effective and used in real time monitoring. The frequent multi-pattern capturing algorithm has high detection rate. Finally find the percentage for detection rate and false alarm rate.

[5] A Survey on Intrusion Detection using Data Mining Techniques - R. Venkatesan, et al.[5]

This paper describes the overview of the intrusion detection system and its each technique. The authors discuss pros and cons of anomaly detection and misuse detection. By combining these two categories and data mining approaches, then include the Apriori association rule algorithm for calculating the confidence levels. Apriori algorithm employs an iterative approach known as a level wise search, where k-item sets are used to explore (k + l)- item sets [5].

[6] A Review of Intrusion Detection System in Computer Networks - Abhilasha A Sayar, et.al.[6]

In this paper the author discuss about the classification of Intrusion detection system, advantageous and disadvantageous and its types. In this the IDS uses the artificial intelligence, fuzzy logic and neural network. The techniques are used to detect the intrusions in the images. For example, in military the original information's are changed into images and then send to another location. By using the artificial intelligence with IDS the user can easily identify the unknown attacks. This paper is useful for beginners to study the basic concepts of Intrusion detection system and also detect all kind of images.

VIII. 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.

A. Feature Selection

Feature Selection ”Feature selection, also known as subset selection or variable selection, is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm. Feature selection is necessary either because it is computationally infeasible to use all available features, or because of problems of estimation when limited data samples are present Feature selection from the available data is vital to the effectiveness of the methods employed. Researchers apply various analysis procedures to the accumulated data, in order to select the set of features that they think maximizes the effectiveness of their data mining techniques.

B. Machine Learning

Machine Learning[8] 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 learning patterns with no a priori knowledge of what those patterns may be. Clustering and Classification are probably the two most popular machine learning problems. Techniques that address both of these problems have been applied to IDSs.

1) Classification 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).

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.

C. Statistical Techniques

Statistical techniques, also known as ”top-down” learning, are employed when we have some idea as to the relationship were looking for and can employ mathematics to aid our search. Three basic classes of statistical techniques are linear, nonlinear (such as a regression-curve), and decision trees [59]. Statistics also includes more complicated techniques, such as Markov models and Bayes estimators. Statistical patterns can be calculated with respect to different time windows, such as day of the week, day of the month, month of the year, etc. [50], or on a per-host, or per-service basis

1) Hidden Markov Models: Much work has been done or proposed involving Markovian models. For instance, the generalized Markov chain may improve the accuracy of detecting statistical anomalies. Unfortunately, it has been noted that these are complex and time consuming to construct [7], however their use may be more feasible in a high-power off-line environment.

IX. CONCLUSIONS

This paper has presented a survey of the various data mining techniques that have been proposed towards the enhancement of IDSs. We have shown the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated by researchers. Finally, in the last section, we proposed a data mining approach that we feel can contribute significantly in the attempt to create better and more effective Intrusion Detection Systems.

REFERENCE

[1] S.A.Joshi, Varsha S.Pimprale, “Network Intrusion Detection System (NIDS) based on Data Mining”, International Journal of Engineering Science and Innovative Technology, Vol. 2, No. 1, January 2013, ISSN. 2319-5967.

[2] Sushil Kumar Chaturvedi, Prof. Vineet Richariya. Prof. Nirupama Tiwari, “Anomaly Detection in Network using Data mining Techniques”, International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 5, May 2012, ISSN. 2250-2459.

[3] Omprakash Chandrakar, Rekha Singh, Dr. Lal Bihari Barik, "Application of Genetic Algorithm in Intrusion Detection System", International Institute for Science, Technology and Education, Vol. 4, No. 1, 2014, ISSN. 2224-5774.

[4] A.R. Jakhale, G.A. Patil, "Anomaly Detection System by Mining Frequent Pattern using Data Mining Algorithm from Network Flow", International Journal of Engineering Research and Technology, Vol. 3, No.1, January 2014, ISSN. 2278-0181.

[5] R. Venkatesan, Dr. R. Ganesan, Dr. A. Arul Lawrence Selvakumar., "A Survey on Intrusion Detection using Data Mining Techniques", International Journal of Computers and Distributed Systems, Vol. 2, No. 1, December 2012, ISSN. 2278-5183.

[6] Abhilasha A Sayar, Sunil. N. Pawar, Vrushali Mane., "A Review of Intrusion Detection System in Computer Network", International Journal of Computer Science and Mobile Computing, Vol. 3, No. 2, February 2014, pp. 700 - 703.

[7] Weiming Hu, Jun Gao, Yanguo Wang, Ou Wu, and Stephen Maybank (2013) Online Adaboost-Based Parameterized Methods for Dynamic Distributed Network Intrusion Detection, IEEE Transactions on Cybernetics. Luigi Coppolino, Salvatore D'Antonio, Alessia

[8] Garofalo, Luigi Romano (2013) Applying Data Mining Techniques to Intrusion detection in Wireless Sensor networks, Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing.

[9] T. Subbulakshmi, Ms. A. Farah Afroze (2013) Multiple Learning based Classifiers using Layered Approach and Feature Selection for Attack Detection, IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN).

[10] Vikas Sharma, Aditi Nema (2013) Innovative Genetic approaches For Intrusion Detection by Using Decision Tree, International Conference on Communication Systems and Network Technologies.