ENHANCING IDS-BASED SECURITY THROUGH DATA MINING

Dr. Savyasachi Assistant Professor, Department of Information Technology L.N. Mishra College of Business Management, Muzaffarpur,Bihar

Abstract – How to efficiently create automatic intrusion rules from acquired raw network data after separating attack patterns and normal data patterns from a huge amount of network data is a key issue in intrusion detection. To do this, a variety of data mining techniques are used, including classification, clustering, association rule mining, etc. Examples of IDS misuse detection models based on data mining include JAM (Java Agents for Meta-learning), MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection), and Automated Discovery of Concise Predictive Rules for Intrusion Detection.

Ant clustering technique in data mining is a novel approach which uses Ants technique to find the relevant information and put them in various clusters. Since several Ants work in parallel therefore the processing speed of the system is high and in case of large data sets it is worth using Ant clustering to apply.

This paper proposes to perform mining on the data collected from the IDS to enhance the speed of detection of intrusion with automatic detection using specific attributes of the intrusions.

Various phases of the proposed work perform data collection, cleaning, clustering, detection and alarming system etc.

Keywords: *Intrusion detection, Probability density function, Genetic Network programming, Data Mining, Ant Clustering.*

1. INTRODUCTION

Intrusion Detection System

A series of acts known as intrusion aim to jeopardize the availability, confidentiality, or integrity of any resource on a computing platform.

To protect data integrity, confidentiality, and system availability against attacks, intrusion detection systems (IDS) are utilized.

A hardware and software device known as an intrusion detection system (IDS) is used to identify network intrusions. IDS can keep an eye on every network activity and as a result, can spot signals of intrusions.

IDS's primary goal is to notify the system administrator of any suspicious activity. An important problem in intrusion detection is how effectively can separate the attack patterns and normal data patterns from a large number of network data and how effectively generate automatic intrusion rules after collected raw network data. Steps in IDS are:

- Data Collection
- Feature Selection
- Analysis
- Action

Intrusion is a set of actions that attempt to compromise the integrity, confidentiality, or availability of any resource on a computing platform. An intrusion detection system (IDS) is a combination of hardware and software that detect intrusions in the network. IDS can monitor all the network activities and hence can detect the signs of intrusions. The main objective of IDS is to alarm the system administrator that any suspicious activity happened. There are two types of Intrusion detection techniques:

- Anomaly Detection: Detecting malicious activities based on deviations from the normal behavior are considered as attacks. Although it can detect unknown intrusions, rate of missing report is low.
- Misuse Detection: Detecting intrusions based on a pattern for the malicious activity. It can be very helpful for known attack patterns. Also rate of missing report is high.

One disadvantage of Misuse Detection over Anomaly Detection is that it can only detect intrusions which contain known patterns of attack.

Working of Intrusion detection systems

Authors have presented a four step approach for the generalized working of IDS

- Data collection: - It involves collecting network traffic using particular software and thus helps to get the information about the traffic like types of packets, hosts and protocol details.
- Feature Selection: - The collected data is substantially large because of the huge network traffic; we generate feature vectors that contain only necessary information. In network-based intrusion detection, it can be IP header information, which consists of source and destination IP address, packet type, layer 4 protocol type and other flags.
- Analysis: - The collected data is analyzed in this step to determine whether data is anomalous or not. Here we use various methods for detecting intrusions.
- Action :- IDS alarm the system administrator that an attack has happened and it tells about the nature of the attack.IDS also participate in controlling the attacks by closing the network port or killing the processes.

Data Mining

Data mining is used to clean, classify and examine large amount of network data. Since a large volume of network traffic that requires processing, we use data mining techniques. Different Data Mining techniques such as clustering, classification and association rules are proving to be useful for analyzing network traffic.

Data Mining is used in variety of applications that requires data analysis. Now a day's data mining techniques plays an important role in intrusion detection systems. Different data mining techniques like Classification, Clustering and Association rules are frequently used to acquire information about intrusions by observing network data. This section describes different data mining techniques that help in detecting intrusions.

Classification

Classification is a form of data analysis which takes each instance of a dataset and assigns it to a particular class. It extracts models defining important data classes. Such models are called classifiers. A classification based IDS will classify all the network traffic into either normal or malicious. Data classification consists of two steps – learning and classification. A classifier is formed in the learning step and that model is used to predict the class labels for a given data in the classification step. Classification analysis requires that the end-user/analyst know ahead of time how classes are defined. Each record in the dataset already has value for the attribute used to define the classes. The objective of a classifier is not to explore the data to discover different classes, but to find how new records should be arranged into classes. Classification helps us to categorize the data records in a predetermined set .It can be used as attribute to label each record and for distinguishing elements belonging to the normal or malicious class. Different types of classification techniques are decision tree induction, Bayesian networks-nearest neighbour classifier, genetic algorithm and fuzzy logic.

As compared to the clustering technique, classification technique is less efficient in the field of intrusion detection. The main reason for this is the enormous amount of data needed to be collected to use classification. To classify the dataset into normal and abnormal, large amount of data is required to analyze its proximity. Classification method can be useful for both misuse detection and anomaly detection, but it is more commonly used for misuse detection.

Authors have presented a data classification for intrusion detection that can be achieved by the following steps:

1. In order to study about the classification models of the normal and abnormal sequences of system calls, we want to supply it with a training data set, containing pre-labelled normal or abnormal sequences. Different techniques like linear discrimination, decision tree or rule based methods is used to scan the network traces. Then generate a collection of unique sequence of system calls and named it as normal list.
2. Next scan each of the intrusion traces. Find each sequence of system calls in the normal list. If an exact match can be found then label it as normal. Otherwise it is labelled as abnormal.
3. Next ensure that the normal traces consist of all possible normal short sequence of system calls. An intrusion trace contains combination of normal and abnormal sequences of system calls since abnormal sequence only appear in some places.

Clustering

Since the amount of available network data is too large, human labelling is time-consuming, and expensive. Clustering is the process of labelling data and assigning into groups.ie, Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of members from the same cluster are quite similar and members from the different clusters are different from each other. Hence clustering methods can be useful for classifying network data for detecting intrusions. Clustering algorithms can be classified into four groups: partitioning algorithm, hierarchical algorithm, density-based algorithm and grid based algorithm.

Clustering techniques can discovers complex intrusions over a different time period. Clustering is an unsupervised machine learning mechanism for discovering patterns in unlabeled data with many dimensions. Clustering is the collection of patterns based on similarity. Patterns within a cluster are equivalent to each other, but they are different with other clusters. Therefore patterns that are far from any of these clusters indicate that an unusual activity happened. That can be part of a new attack.

Clustering can be applied on both Anomaly detection and Misuse detection.

Authors have presented basic steps involved in identifying intrusion are follows:-

1. Find the largest cluster, which consists of maximum number of instances, and label it as normal.
2. Sort the remaining clusters in an ascending order of their distances to the largest cluster.
3. 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.
4. Label all other clusters as malicious.
5. After clustering, heuristics are used to automatically label each cluster as either normal or malicious. The self labelled clusters are then used to detect attacks in a separate test dataset.

Data Mining & IDS

The data mining technology have the capability of extracting large databases; it is of great importance to use data mining techniques in intrusion detection. Data mining techniques plays an important role in

intrusion detection systems. Different data mining techniques like Classification, Clustering and Association rules are frequently used to acquire information about intrusions by observing network data.
IDS through Data Mining are mainly done on the following two types of classifications:
Anomaly Detection: Detecting malicious activities based on deviations from the normal behavior are considered as attacks. Although it can detect unknown intrusions, rate of missing report is low.
Misuse Detection: Detecting intrusions based on a pattern for the malicious activity. It can be very helpful for known attack patterns. Also rate of missing report is high.
IDS may use either anomaly based approach or misuse based approach. Traditional IDS were making use of misuse based approach. The drawback of misuse based approach is that it cannot detect new type of attacks. Hence anomaly based intrusion detection were used and it is capable of detecting unknown attacks also. Anomaly based intrusion detection makes extensive use of data mining because of the advantage it provides.

ANT CLUSTERING

Data clustering, or just clustering, is an explorative task that seeks to identify groups of similar objects based on the values of their attributes. Clustering works on the inherent characteristics of the data and attempts to discover distinct boundaries to divide the data set into meaningful partitions. Deneubourg et al. proposed a basic model which generalized the clustering behavior of ants into two simple actions:
Picking up an isolated item, and
Dropping the item where more similar items are present.
Assuming the ants or agents can only handle one item at a time and only one type of item exists in the environment, they defined each action in terms of a probabilistic function.
Other similar work includes the AntCluss clustering algorithm, which is a combination of an ant colony with the partitional K-Means algorithm. The ant colony of AntClass differs from Lumer & Faieta’s model as ants are allowed to carry more than a single object at a time, have local memory and other heterogeneous features.
Ant-based techniques, in the computer sciences, are designed for those who take biological inspirations on the behavior of these social insects. Data-clustering techniques are classification algorithms that have a wide range of applications, from Biology to Image processing and Data presentation. Since real life ants do perform clustering and sorting of objects among their many activities, we expect that a study of ant colonies can provide new insights for clustering techniques.
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 share some common trait. Clustering is used as a data processing technique in many different areas of application, such as bioinformatics, data mining, image analysis, etc.

2. EXISTING SYSTEM

Authors propose that every enterprise wants to protect their data from both the internal and external attackers. In this initiative firewall, encryption, and authentication serve as the first line of defence. And Intrusion Detection serves as the second line of defence. IDS may use either anomaly based approach or misuse based approach. Traditional IDS were making use of misuse based approach. The drawback of misuse based approach is that it cannot detect new type of attacks. Hence anomaly based intrusion detection were used and it is capable of detecting unknown attacks also. Anomaly based intrusion detection makes extensive use of data mining because of the advantage it provides. In this paper we try to explore the features of intrusion detection based on data mining. [1]

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.

In their work, authors says that nowadays, as information system plays critical part in the internet, the importance of secure networks is tremendously increased. Intrusion Detection System (IDS) is used to preserve the data integrity, confidentiality and system availability from attacks. Data mining is used to clean, classify and examine large amount of network data. Since a large volume of network traffic that requires processing, we use data mining techniques. Different Data Mining techniques such as clustering, classification and association rules are proving to be useful for analyzing network traffic. This paper presents the survey on data mining techniques applied on intrusion detection systems for the effective identification of both known and unknown patterns of attacks, thereby helping the users to develop secure information systems. [2]

In this paper, we describe different data mining technique applied for detecting intrusions. This paper provides the details of two types of intrusion detection and general working principle of IDS.Misuse detection techniques are not sufficient for identifying unknown attacks. For detecting unknown intrusions, we need to go for anomaly detection. Also this paper presents the main concepts of data mining process and the system design for data mining based intrusion detection pattern. Different Data mining techniques like classification, clustering and association rule are very helpful in analyzing the network data. Since large amount of network traffic needs to be collected for intrusion detection, clustering is more suitable than classification in the domain of intrusion detection. Data mining technology helps to understand normal behavior inside the data and use this knowledge for detecting unknown intrusions. [2]

The paper[3] focuses on an improved FP-Growth algorithm. Pre-processing of data mining can increase efficiency on searching the common prefix of node and reduce the time complexity of building FP-tree. Based on the improved FPGrowth algorithm and other data mining techniques, an intrusion detection model is carried out. The experimental results demonstrate effectiveness of the improved algorithm and feasibility of intrusion detection model.

The application of data mining techniques to IDS is one important direction for future development of intrusion detection. Selecting appropriate data mining algorithms and designing IDS model are effective measures in order to improve system detection performance. In this paper, Kmeans and DSFP-Growth algorithm are used to construct the intrusion detection model. The performance study shows that it can reduce runtime, increase detection rate and false positive rate. This model is efficiently in large databases. In order to reduce false negative rate and improve detection accuracy of the system model, the further research should be focused on improving the structure of intrusion detection model and adjusting experimental parameters. [3]

In response of the fact that traditional intrusion detection systems are not able to fulfill the requirements for specific network security, such as fast processing speed, stronger defense capability, and higher real-time performance, a model of network security defense is built on the integration of data stream mining and intrusion detection; and, a data stream clustering algorithm is designed for mining in the model. Through analysis and simulation, the model turns out to be higher in detection rate and lower in false-alarming or false negative rate, thus achieving a better result. [4]

The data-mining-based intrusion detection system demonstrates a good detection performance, high in detection rate, low in false alarm or false negative rate, and better at recognition of specific intrusion types. [4]

In conventional network security simply relies on mathematical algorithms and low counter measures to taken to prevent intrusion detection system, although most of this approaches in terms of theoretically challenged to implement. Therefore, a variety of algorithms have been committed to this challenge. Instead of generating large number of rules the evolution optimization techniques like Genetic Network Programming (GNP) can be used .The GNP is based on directed graph, In this paper the security issues related to deploy a data mining-based IDS in a real time environment is focused upon. We generalize the problem of GNP with association rule mining and propose a fuzzy weighted association rule mining with GNP framework suitable for both continuous and discrete attributes. Our proposal follows an Apriori algorithm based fuzzy WAR and GNP and avoids pre and post processing thus eliminating the extra steps during rules generation. This method can sufficient to evaluate misuse and anomaly detection. Experiments on KDD99Cup and DARPA98 data show the high detection rate and accuracy compared with other conventional method. [5]

The proposed method, that is, Fuzzy weighted association rule mining algorithm based on GNP, combines genetic algorithm and probabilistic classification in order to extract more important rules from the database. In addition, the classification method is based on the probability distribution of the average matching degree between data and different class rules. As a result, simulations show higher DR, Accuracy and lower PFR, NFR, which means that Fuzzy weighted association rule mining algorithm based on GNP has better performance than the conventional class association rule mining.

3. PROBLEM STATEMENT

From the various works studied it is found that there are a large number of tools available for IDS implementation and lot of research is undergoing in the field.

Lot of scope has been seen in the field of IDS through Data Mining and also the advantages of IDS implementation using Data Mining are encouraging to work further in this topic.

An important problem in intrusion detection is how effectively can separate the attack patterns and normal data patterns from a large number of network data and how effectively generate automatic intrusion rules after collected raw network data. To accomplish this, various data mining techniques are used such as classification, clustering, association rule mining etc. Examples for Data Mining based Misuse detection model of IDS are JAM (Java Agents for Meta-learning), MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection), and Automated Discovery of Concise Predictive Rules for Intrusion Detection.

Examples for Data Mining based Anomaly detection model for IDS are MINDS and EBays. Examples for Data Mining based both Anomaly and Misuse detection model for IDS are IIDS (Intelligent Intrusion Detection System Architecture) and RIDS-100 (Rising Intrusion Detection System).

4. PROPOSED ALGORITHM

Ant clustering technique in data mining is a novel approach which uses Ants technique to find the relevant information and put them in various clusters. Since several Ants work in parallel therefore the processing speed of the system is high and in case of large data sets it is worth using Ant clustering to apply. The algorithm proposed is as follows:

Data Collection Phase: In this phase, C# will be applied to collect the incoming packets in a machine and relevant details of the packets shall be extracted to create a data collection. This phase will prepare the data as required for intrusion detection.

Data Cleaning: Lot of data shall be collected and therefore it is cleaned in this phase using the intrusion characteristics such as source and destination IP address, ports used, frequency of incoming packets etc.

Ant Clustering: In this phase ant clustering shall be applied to the cleaned data for clustering data in various groups for identifying the intrusions.

Decision Making & Result Generation: In this phase clustered data shall be used to make the decisions related with the intrusion and various results shall be drawn from the decision done.

Alarm System: In this phase users / administrators shall be informed about the intrusions and non intrusions and various reports shall be generated for showing the improvements using the application of the any clustering based data mining.

The studies of various algorithms of data mining have been done along with the study of the intrusion detection systems and the characteristics of the intrusions. Since Ant Clustering has been seen to be providing the fastest processing therefore it is chosen as the data mining technique for the proposed work.

Tools decided for implementation are C# and MS SQL Server and they are being studied so that final implementation can be done.

The project work implementation will have the following screens:

- Creating Various Screens for user interfacing
- Implementation of data collector utility using C#
- Applying cleaning of the collected data using characteristics
- Applying Ant Clustering to cluster the cleaned data
- Decision making related with the clusters for intrusions and non-intrusion data
- Generating an alarm system for the users/administrator
- Generating reports & graphs for the proposed work
- Comparing the existing system with the proposed system for improvements.

5. CONCLUSION

The various research papers have been studied and decided the field. After reading and through guidance decided the final topic for implementation as dissertation work.

The studies of various algorithms of data mining have been done along with the study of the intrusion detection systems and the characteristics of the intrusions. Since Ant Clustering has been seen to be providing the fastest processing therefore it is chosen as the data mining technique for the proposed work.

Tools decided for implementation are C# and MS SQL Server and they are being studied so that final implementation can be done.

REFERENCES

[1] M. Moorthy, Dr. S. Sathiyabama, “A Study of Intrusion Detection using Data Mining“ IEEE- International Conference On Advances In Engineering, Science And Management (ICAESM -2022) March 30, 31, ISBN: 978-81-909042-2-3 ©2012 IEEE

[2] Deepthy K Denatious & Anita John, “Survey on Data Mining Techniques to Enhance Intrusion Detection“ 2012 International Conference on Computer Communication and Informatics (ICCCI -2022), Jan. 10 – 12, 2012, Coimbatore, INDIA, 978-1-4577-1583-9/ 12/ $26.00 © 2012 IEEE

[3] LI Yin–huan, “Design of Intrusion Detection Model Based on Data Mining Technology” 2012 International Conference on Industrial Control and Electronics Engineering 978-0-7695-4792-3/12 © 2012 IEEE DOI 10.1109/ICICEE.2012.156

[4] Zhu Lin, Zhu Can Shi, “Research into the Network Security Model Blended of Data Stream Mining and Intrusion Detection System” The 7th International Conference on Computer Science & Education (ICCSE 2012) July 14-17, 2022. Melbourne, Australia, 978-1-4673-0242-5/12 ©2012 IEEE

[5] P. Prasenna, A.V.T RaghavRamana, R. KrishnaKumar, A. Devanbu, “Network Programming And Mining Classifier For Intrusion Detection Using Probability Classification“ Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2022, 978-1-4673-1039-0/12 ©2022 IEEE

[6] Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Associations between Sets ofltems in Massive Databases. In Proceedings of the ACM-ST OMOD I991ntemational Conferencing Management of Data, pages 207-216.

[7] Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases, pages 487{499}.

[8] Berry, M. 1. A. and Lino-, O. (1997). Data Mining Techniques. John Wiley and Sons, Inc.

[9] Biswanath Mukherjee, L.Todd Heberlein, Karl .Levitt, "Network Intrusion etection",IEEE, June 1994.

[10] Barbara, D., Couto, 1., Jajodia, S., Popyack, L., And Wu,N., ADAM: Testbed for Exploring the Use of Data Miningin Intrusion Detection, ACM SI OMOD Record, 30(4), 200I,pp. 15-24.

[11] Chittur, A., "Model generation for an intrusion detection system using genetic algorithms", High School Honors Thesis, Ossining High School. In cooperation with Columbia Univ, 2001