Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

A Comparability Study of Intrusion Detection System using Data Mining Techniques

K. Raja[1], Dr. M. Lilly Florence[2], and Dr. D. Swamydoss[3]
Assistant Professor[1], Professor[2], Professor & HOD[3]
Department of MCA, Adhiyamaan College of Engineering, Hosur, TamilNadu, India.

ABSTRACT: The Main objective of this paper is to avoid the intrusion using data mining techniques with help of multi agents. Data mining is a discovery process that allows users to understand the substance of and the relationships between, their data. Data mining uncovers patterns and rends in the contents of this information. Intrusion detection systems have been used along with the data mining techniques to detect intrusions. In this work we aim to use data mining techniques including classification tree and support vector machines for intrusion detection. To meet the challenges of both efficient learning (mining) and real-time detection, we propose an agent based architecture for intrusion detection systems where the learning agents continuously compute and provide the updated (detection) models to the detection agents. Intrusion detection is therefore needed as another wall to protect computer systems.

Keywords: *Data Mining, Intrusion Detection System (IDS), Preprocessing, Decision Tree, Clustering Techniques, Intrusion Detection Technique*.

I. INTRODUCTION

Intrusion detection can be defined as identifying individuals who are using a computer system without authorization and those who have legitimate access to the system but are abusing their privileges. Intrusion Detection system (IDS) prepare for and deal with attacks by collecting information from a variety of system and network sources, then analyzing the symptoms of security problems. A secure network must provide the following:

Data confidentiality: Data that are being transferred through the network should be accessible only to those that have been properly authorized.

Data integrity: Data should maintain their integrity from the moment they are transmitted to the moment they are actually received. No corruption or data loss is accepted either from random events or malicious activity.

Data availability: The network should be resilient to Denial of Service attacks [1].

II. DATA MINING

Data mining refers to extracting or "mining" knowledge from large amounts of data "Knowledge mining," a shorter term, may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material (Figure 1). Thus, such a misnomer that carries both "data" and "mining" became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.

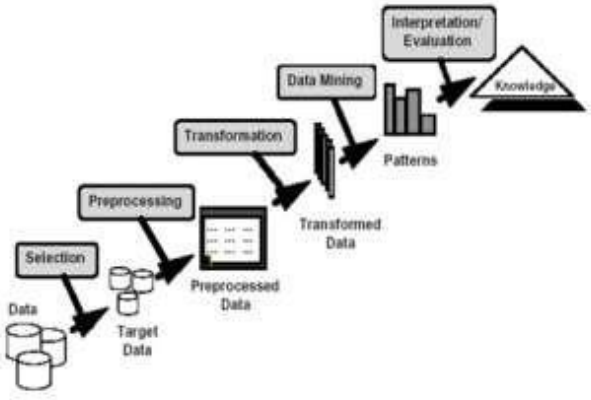

Figure 1: Data Mining Process

Data mining or knowledge discovery in databases, as it is also known is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes, and detecting **anomalies**.

III. INTRUSION DETECTION SYSTEM (IDS)

An intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a Management Station. Some systems may attempt to stop an intrusion attempt but this is neither required nor expected of a monitoring system. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, and reporting attempts [3].

IV. PREPOCESSING

Preprocessing is the data cleaning stage where unnecessary information is removed. For example, it is unnecessary to

note the sex of a patient when studying pregnancy! When the data is drawn from several sources, it is possible that the same information represented in different sources in different formats. This stage reconfigures the data to ensure a consistent format, as there is a possibility of inconsistent formats [4].

A. Data Cleaning

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it [4].

B. Data Integration

This would involve integrating multiple databases, data cubes, or files, that is, data integration. For example, the attribute for customer identification may be referred to as customer_id in one data store and cus_id in another. Naming inconsistencies may also occur for attribute values. For example, the same first name could be registered as "Bill" in one database, but "William" in another, and "B." in the third [4].

C. Data Transformation

It would be useful for your analysis to obtain aggregate information as to the sales per customer region something that is not part of any precomputed data cube in your data warehouse. You soon realize that data transformation operations, such as normalization and aggregation, are additional data preprocessing procedures that would contribute toward the success of the mining process [4].

D. Data Reduction

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. There are a number of strategies for data reduction. These include data aggregation (e.g., building a data cube), attribute subset selection (e.g., removing irrelevant attributes through correlation analysis), dimensionality reduction (e.g., using encoding schemes such as minimum length encoding or wavelets), and numerosity reduction (e.g., "replacing" the data by alternative, smaller representations such as clusters or parametric models) [4]

V. CLASSIFICATION OF CLUSTERING TECHNIQUES

Clustering algorithms are broadly classified as

i) Hierarchical clustering.
ii) Partitional clustering.

The following figure shows the different techniques of clustering,

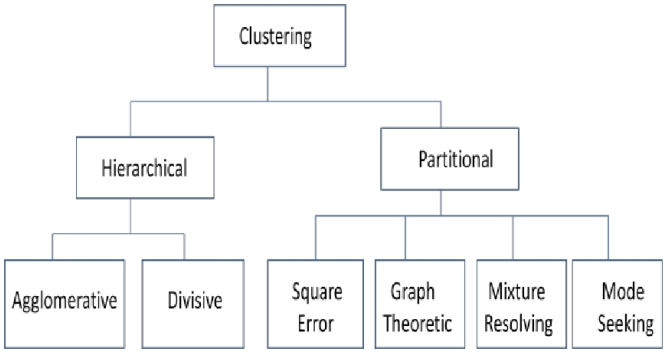

Figure 3: Classification of Clustering Technique

A. Hierarchical Clustering

Hierarchical clustering, clusters the given image, based on the concept of pixel being more closely related to nearby pixels than the pixels which are farther i.e. These algorithms groups the pixels into cluster based on their distances. Hierarchical clustering represents data into a tree structure form. In which the whole data set is represented by root node and the individual data points are represented by leaf node. The intermediate nodes in a tree structure represent the similarity among the pixel data points. In this clustering technique numbers of algorithms are proposed based on the method by which distances are computed. Along with the distance function the linkage criterion is also an important factor in hierarchical clustering [5].

1. Agglomerative

A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds [5].

2. Divisive

The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds [5].

B. Partitional Clustering

Partitional Clustering algorithms separate the pixels or data points into number of partitions. These partitions are referred as clusters. The partitional clustering organizes data into single partition instead of representing data into nested structure like hierarchical clustering. The partitional clustering is classified as square error clustering, Graph theoretic clustering, mixture resolving clustering and mode seeking clustering [6]

1. Square Error Clustering

The most frequently used convergence criterion in partitional criterion is squared error algorithms. The main advantage of

square error algorithms is that it works well with isolated and compact clusters [6].

2. Graph Theoretic Clustering

Graph Theoretic Clustering algorithm is divisive clustering algorithm. This algorithm first forms the minimal spanning tree (MST) for the given data. Then it obtained clusters by deleting edges of largest length from the minimal spanning tree structure [6].

3. Mixture-Resolving Clustering

Clustering can be understood by studying density distribution functions. In the mixture resolving algorithm the parametric distribution function like Gaussian distribution are used and the vectors of component density are form. These vectors are grouped together iteratively based on maximum likelihood estimation to form the clusters[6].

4. Mode-Seeking Clustering

In non-parametric technique the algorithms are developed inspired by the Parzen window approach. It forms clusters by creating bins with large counts in multidimensional histogram of the input mixture patterns. This approach considered as mode seeking approach [6].

VI. INTRUSION DETECTION TECHNIQUE

Intrusion detection systems (IDSs) are usually deployed along with other preventive security mechanisms such as access control and authentication, as a second line of defense that protects information systems.. First, many traditional systems and applications were developed without security in mind (For example, a system may be perfectly secure when it is isolated but become vulnerable when it is connected to the Internet.). Second, due to the limitations of information security and software engineering practice, computer systems and applications may have design flaws or bugs that could be used by an intruder to attack the systems or applications. As a result, certain preventive mechanisms (e.g., firewalls) may not be as effective as expected [7].

A. Anomaly Detection

Anomaly-Based Intrusion Detection System is a system for detecting computer intrusions and misuse by monitoring system activity and classifying it as either normal or anomalous. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection). Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model [7].

1. Statistical anomaly-based IDS

A statistical anomaly-based IDS determines normal network activity like what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous(not normal).

2. Signature-based IDS

Signature based IDS monitors' packets in the Network and compares with pre-configured and pre-determined attack patterns known as signatures.

B. Misuse Detection

Misuse detection is considered complementary to anomaly detection. The rationale is that known attack patterns can be detected more effectively and efficiently by using explicit knowledge of them. Misuse detection is an approach in detecting attacks. In misuse detection approach, we define abnormal system behavior at first, and then define any other behavior, as normal behavior. It stands against anomaly detection approach which utilizes the reverse approach [7].

C. Intrusion Detection in Distributed Systems

Early distributed IDSs collect audit data in a distributed manner but analyze the data in a centralized place, for example DIDS (Snapp et al., 1991) and ASAX(Mounji, Charlier, Zampunieris, & Habra, 1995). To scale up to large distributed systems, these systems place IDS components in various places in a distributed system. Each of these components receives audit data or alerts from a limited number of sources (e.g., hosts or other IDS components), so the system is not overwhelmed by large amounts of audit data [7].

CONCLUSION

Data mining is the process of analyzing data from different perspectives and summarizing into useful information. Data mining can be divided into four types: association analysis, sequence analysis, classification analysis and cluster analysis. In this paper data mining techniques to detect the intrusion in multi-agents system. Intrusion detection continues to be an active research field. Even after 20 years of research, the intrusion detection community still faces several difficult problems. How to detect unknown patterns of attacks without generating too many false alerts remains an unresolved problem, although recently, several results have shown there is a potential resolution to this problem. Beside this it will have single point system administration and maintenance which makes it user friendly. It will also add security to a system as well as networks. Through this proposed system it is easy to mention an IP address.

References

[1] An Introduction to Intrusion Detection Systems"By Paul Innella and Oba McMillan, Tetrad Digital Integrity, LLC http://www.securityfocusonline.com.

[2] Adriaans, Pieter, Data mining, Delhi: Pearson Education. Asia, 1996.

[3] Scarfone, Karen; Mell, Peter (February 2007). "Guide to Intrusion Detection and Prevention Systems (IDPS)".Computer Security Resource Center (National Institute of Standards and Technology) (800–94). Retrieved 1 January 2010.

[4] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2002.

[5] Juntao Wang, Xiaolong Su, An improved K-Means clustering algorithm , IEEE proceeding, pp. 44-46, 2011.

[6] Ashwini Gulhane, Prashant L. Paikrao, D. S. Chaudhari "A Review of Image Data Clustering Techniques, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[7] Axelsson, S. (1999). Research in intrusion-detection systems: A survey. Technical report TR 98-17. Göteborg, Sweden: Department of Computer Engineering, Chalmers University of Technology.