Leverage Data Mining Techniques in Intrusion detection

S. Neelima¹, N. Satyanarayana² and P. Krishna Murthy³

¹Department of CSE, Research Scholar, JNTUH, Hyderabad, India.
²Department of CSE, Nagole Institute of Technology and Science, Hyderabad, India.
³Principal, Swarna Bharathi Institute of Science and Technology, Khammam, AP, India

Abstract—Data mining techniques have been successfully applied in many fields including marketing, manufacturing, process control, fraud detection and network management. In recent years, data mining-based Intrusion Detection Systems (IDS) have demonstrated high accuracy, good generalization to novel types of intrusion, and robust behavior in a changing environment. In data mining based intrusion detection system one should have knowledge about the particular domain in relation to intrusion detection so as to efficiently extract relative rule from huge amounts of records. This paper presents a review of the literature in connection with different data mining techniques for intrusion detection.

Keywords — Data Mining, Intrusion Detection, Data Mining Techniques.

I. INTRODUCTION

The area of data mining is gaining importance due to availability of large volume of data, easily collected and stored in electronic format. The large amount of data gathered from the different sources may contain personal and sensitive data. During the process of data mining an important aspect is to provide the privacy and security when mining of data.

From past decade data mining is paying more attention because of its enormous area of applications. For the security point of view, data mining may be helpful in confronting various types of security attacks to the universe. Data mining technology focuses on discovery of general or statistically significant patterns [1].

In past year’s research have started looking into the possibility of using data mining techniques in the emerging field of information security especially in the challenging problem of intrusion detection. Intrusion is commonly defined as a set of actions that attempt to violate the integrity, confidentiality or availability of a system. Intrusion detection is the process of finding important events occurring in a computer system and analyzing them for possible presence of intrusion. So, it is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems [2]. This paper focuses on the intrusion detection system and data mining techniques in intrusion detection.

II. DATA MINING

Data mining (DM) also called Knowledge-Discovery and is one of the hot topic in the field of knowledge extraction from database. Data mining is used to automatically learn patterns from large quantities of data. Mining can efficiently discover useful and interesting rules from large collection of data. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining is disciplines works to finds the major relations between collections of data and enables to discover a new and anomalies behavior [3]. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. A primary reason for using data mining is to assist in the analysis of collections of observations of behavior.

III. INTRUSION DETECTION

Intrusion activities of computer systems are increasing due to the commercialization of the internet & local networks. An IDS, collects the information and analyzing it for uncommon or unexpected events. ID is the process of monitoring and analyzing the events which occurred in a computer system in order to detect signs of security problems. Figure 1 shows a taxonomy of Intrusion Detection Systems [4].

![Figure 1. Taxonomy of Intrusion Detection System](image_url)

A. IDS Based on Technique

Anomaly Detection: In anomaly detection, the system defines the expected behavior of the network (or profile) in advance.
Any significant deviations from this expected behavior are then reported as possible attacks. Such deviations are not necessarily actual attacks. They may simply be new network behavior that needs to be added to the profile. The primary advantage of anomaly-based detection is the ability to detect novel attacks for which signatures have not been defined [1]. This technique is based on the detection of traffic anomalies. The deviation of the monitored traffic from the normal profile is measured. Various different implementations of this technique have been proposed, based on the metrics used for measuring traffic profile deviation [11].

**Misuse/Signature Detection:** Misuse detection finds intrusions by looking for activity corresponding to known techniques for intrusion. This generally involves the monitoring of network traffic in search of direct matches to known patterns of attack (called signatures). This is essentially a rule-based approach [5]. 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.

**Hybrid Detection:** This technique uses combination of both Anomaly detection and Misuse detection [6].

### B. IDS Based on Data Target

**Host Based IDS:** Its data come from the records of various host activities, including audit record of operation system, system logs, application programs information, and so on.

**Network-Based IDS:** Its data is mainly collected network generic stream going through network segments, such as: Internet packets.

**Hybrid IDS:** It is combination of hosed based IDS and Network based IDS [6].

### IV. USAGE OF DATA MINING IN INTRUSION DETECTION

The goal of intrusion detection is to detect security violations in information systems. Intrusion detection is a passive approach to security as it monitors information systems and raises alarms when security violations are detected[7]. In the past five years, a growing number of research projects have applied data mining to intrusion detection. The term data mining is used to describe the process of extracting useful information from the large databases. Data mining analyses the observed sets to discover the unknown relation and sum up the results of data analysis to make the owner of data to understand.

Hence data mining problems are considered as a data analysis problem. Data mining framework automatically detect patterns in our data set and use these patterns to find a set of malicious binaries i.e., Data mining techniques can detect patterns in large amount of data, such as byte code and use these patterns to detect future instances in similar data.

Jun S et al [8] presented a system design based on Data Mining intrusion detection pattern which is illustrated in Figure 2.

![Data Mining Based Intrusion Detection Pattern](image)

**Figure 2. Data Mining Based Intrusion Detection Pattern**

### V. DATA MINING TECHNIQUES IN INTRUSION DETECTION

Data mining techniques play an important role in intrusion detection systems. 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 [9].

#### B. 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 [2].

C. 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 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 [9].

D. 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 [9].

E. 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 [2]. 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.

Data mining techniques are widely used because of their capability to drastically improve the performance and usability of intrusion detection systems. Different data mining techniques like classification, clustering and association rule mining are very helpful in analyzing the network data.

Acknowledgement

The authors would like to thank Anonymous Reviewers for their valuable suggestions and comments. This paper has greatly benefited from their Efforts.

REFERENCES

[1] Chandrashekar Azad , Vijay Kumar Jha ”Data Mining in Intrusion Detection: A Comparative Study of Methods, Types and Data Sets”IJ. Information Technology and Computer Science, 2013, 08, 75-90

