A Survey on Intrusion Detection using Data Mining Techniques

R. Venkatesan
Research Scholar/CS,
CMJ University,
Shillong, Meghalaya.

Dr. R. Ganesan
Professor/ E& I
Noorul Islam University
Kumaracoil, Thuckalay,

Dr. A. Arul Lawrence Selvakumar
Professor & Head, Dept of CSE,
Rajiv Gandhi Institute of Technology,
Bangalore, Karnataka.

ABSTRACT

Intrusions are the activities that violate the security policy of system. Intrusion Detection is the process used to identify intrusions. Network security is to be considered as a major issue in recent years, since the computer network keeps on extending dramatically. Information Systems and Networks are subject to electronic attacks and the possibilities of intrusion are very high. An Intrusion Detection System (IDS) is a system for detecting intrusions and reporting to the authority or to the network administration. Data mining techniques have been successfully applied in many fields like Network Management, Education, Biology, Marketing, Manufacturing, Process control and Fraud Detection. Data Mining for IDS is the technique which can be used mainly to identify unknown attacks and to raise alarms when security violations are detected. The purpose of this survey paper is to describe the methods/techniques which are being used for Intrusion Detection based on Data mining concepts.

KEYWORDS: Data Mining, Intrusion Detection System (IDS), Network Security, Misuse Detection, Anomaly Detection, Classification, Clustering

1. INTRODUCTION

In recent years many researchers are focusing to use Data Mining concepts for Intrusion Detection. Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the huge data. In the other hand intrusions in an information system are the activities that violate the security policy of the system, and intrusion detection is the process used to identify intrusions. This is survey paper; we are going to study the Data Mining approaches which are being followed to detect intrusion in a network.

1.1 DATA MINING

Data mining is the process of discovering interesting patterns (or knowledge) from large amounts of data. The data sources can include databases, data warehouses, the Web, any other information repositories, or data that are streamed into the system (dynamically). Data can be associated with classes or concepts that can be described in summarized, concise, and yet precise, terms. Such descriptions of a concept or class are called class/concept descriptions. These descriptions can be derived via Data Characterization and Data Discrimination.

The Knowledge Discovery in Databases (KDD) Process is used to denote the process of extracting useful knowledge from large data sets. The KDD process involves a number of steps and is often interactive, iterative and user-driven decision making rules [1]: Data mining is the most vital step in the KDD process, and it applies data mining techniques to extract patterns from the data.

- Know the application domain: to understand the back ground of the knowledge and to specify the goal.
- Data Collection: includes creating a target dataset which is relevant to the analysis
- Data Mining: applying an appropriate algorithm to extract useful information using techniques.
- Data Interpretation: to understand the discovered patterns and to confirm the goal is achieved.
- Knowledge Representation: the final stage of representing the discovered knowledge.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks and it can be classified into two categories:

- Descriptive: to characterize the general properties of data in the database
- Predictive: to perform inference on data and to make predictions

1.2 INTRUSION DETECTION

The Intrusion Detection concept was introduced by Anderson J. in 1980 [2]. 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. The IDSs may be classified into (i) Host-base IDSs (ii) Distributed IDSs and (iii) Network-based IDSs.
- **Host-based IDSs** - Host based IDSs examine data held on individual computer that serve as hosts. The network structural design of host based is an agent-based, which means that software resides on each of the hosts that will be governing by the system.
- **Distributed IDSs** - gather audit data from multiple hosts and possibly the network that connects the hosts, aiming at detecting attacks involving multiple hosts.
- **Network-Based IDS** - Uses network traffic as the audit data source, relieving the burden on the hosts that usually provide normal computing services and detect attacks from network.


- **Subjects** refer to the initiators of activity in an information system; they are usually normal users.
- **Objects** are the resources managed by the information system, such as files, commands and devices.
- **Audit records** are those generated by the information system in response to actions performed or attempted by subjects on objects. Examples include user login, command execution, etc.
- **Profiles** are structures that characterize the behavior of subjects with respect to objectives in terms of statistical metrics and models of observed activity.
- **Anomaly records** are indications of abnormal behaviors when they are detected.
- **Activity rules** specify actions to take when some conditions are satisfied, which updates the profiles, detects abnormal behaviors, relate anomalies to suspected intrusions, and produce reports.

## 2. INTRUSION DETECTION TECHNIQUES

The intrusion detection techniques based upon data mining [4], [5] are generally fall into one of two categories: **anomaly detection** and **misuse detection**. The signatures of some attacks are known, whereas other attacks only reflect some deviation from normal patterns.

### 2.1 ANOMALY DETECTION

Anomaly detection attempts to determine whether deviation from an established normal behavior profile can be flagged as an intrusion [6]. Anomaly detection consists of first establishing the normal behavior profiles for users, programs, or other resources of interest in a system, and observing the actual activities as reported in the audit data to ultimately detect any significant deviations from these profiles. Most anomaly detection approaches are statistical in nature.

Anomaly detection may be divided into static and dynamic anomaly detection. A static anomaly detector is based on the assumption that there is a portion of the system being monitored that does not change. The static portion of a system is the code for the system and the constant portion of data depends upon the correct functioning of the system. Dynamic anomaly detection typically operates on audit records or on monitored networked traffic data. An audit record of operating systems does not record all events; they only record events of interest. Strength of anomaly detection is its ability to detect previously unknown attacks.

### 2.2 MISUSE DETECTION

Misuse detection works by searching for the traces or patterns of well-known attacks. Lee et al. [7] designed a signature-based database intrusion detection system (DIDS) which detects intrusions by matching new SQL statements against a known set of transaction fingerprints. 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. This system usually searches for patterns or user behavior that matches known intrusion or scenarios, which are stored as signatures. If a pattern match is found, it signals an event then an alarm is raised. But it is unable to detect new or previously unknown intrusion. Pattern, Data mining, and state transition analysis are some of the approaches of misuse detection. To perform this detection method, each scenario need to described or modeled. Misuse detection is based on extensive knowledge of patterns associated with known attacks provided by human experts.

### 2.3 PROS AND CONS OF ANOMALY AND MISUSE DETECTION

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. 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.

| Table 1: Pros and Cons of Anomaly Detection and Misuse Detection |
|--------------------|-----------------------------|-----------------------------|
| **Technique**       | **Pros**                    | **Cons**                    |
| Anomaly Detection   | Is able to detect unknown attacks based on audits | High false-alarm and limited by training data. |
| Misuse Detection    | Accurately and generate much fewer false alarm | Cannot detect novel or unknown attacks |
2.3 COMBINING MISUSE AND ANOMALY DETECTION
Anomaly detection and misuse detection have major shortcomings that hamper their effectiveness in detecting intrusions. Research can be carried into intrusion detection methodologies which combine the anomaly detection approach and the misuse detection approach. These techniques seek to incorporate the benefits of both the standard approaches to intrusion detection. The combined approach permits a single intrusion detection system to monitor for indications of external and internal attacks. Pattern recognition possesses a distinct advantage over anomaly and misuse detection methods in that it is capable of identifying attacks which may occur over an extended period of time, a series of user sessions, or by multiple attackers working in concert.

2.4 DRAWBACKS OF CURRENT IDS
Intrusion Detection Systems (IDS) have become a standard component in security infrastructures as they allow network administrators to detect any violations. These security violations range from external attackers trying to gain unauthorized access to insiders abusing their access. Current IDS have a number of significant drawbacks [9]:

- **False Positives** – A common complaint is the amount of false an IDS will generate.
- **False Negatives** – In this case, IDS does not create a signature or alarm, when an intrusion is actually happened.
- **Data Overload** – In this aspect, Misuse Detection cannot be related directly, however, it is very important to analyze how much data an analyst can efficiently and effectively analyze.

3. DATA MINING APPROACHES
Data mining generally refers to the process of (automatically) extracting models from large stores of data [10]. The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database. There are several types of algorithms [10] which are particularly related to intrusion detection.

- **Classification**: classifies a data item into one of several pre-defined categories. These algorithms normally output “classifiers”. An ideal application in intrusion detection would be to gather sufficient “normal” and “abnormal” audit data for a user or a program, then apply a Classification algorithm to learn a classifier that can label or predict new unseen audit data as belonging to the normal class or the abnormal class;
- **Link analysis**: determines relationships between fields in the data base records. Correlations of system features in audit data, for example, the correlation between command and argument in the shell command history data of a user, can serve as the basis for constructing normal usage profiles;
- **Sequence analysis**: models sequential patterns. These algorithms can discover what time-based sequences of audit events are frequently occurring together. These frequent event patterns provide guidelines for incorporating temporal and statistical measures into intrusion detection models.

3.1 ASSOCIATION RULE OR DEPENDENCY MINING:
Association analysis is the discovery of association rules showing attribute – value conditions that occur frequently together in a given set of data. Association analysis widely used in transaction data analysis. This approach work on data dependency, in which one item is modify another item refer with this also modify. The concept of Association mining is to find all co-occurences relationship called associations. Association Mining has been used in various domains and many efficient algorithms, extensions and applications have reported. In general, Association analysis has been considered as an unsupervised technique, so it can be applied for KDD task.


**Algorithm:** let A be a set of attributes, and I be a set of values on A, called items. Any subset of I is called an item set. The number of items in an item set is called its length. Let D be a database with n attributes (columns). Define support (X) as the percentage of transactions (records) in D that contain item set X. An association rule is the expression

\[ X \rightarrow Y, \text{[confidence, support]} \]

Here X and Y are item sets, and \( X \cap Y = 0 \). \( \text{support}(X U Y) \) is the support of the rule, and \( \text{support}(X \cup Y) / \text{support}(X) \) is the confidence of the rule.

3.1.2 Apriori Association Rules Algorithm [Agrawal and Srikant, 1994].

Apriori is an algorithm for mining frequent item sets for Boolean association rules [12]. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. Apriori algorithm employs an iterative approach known as a level wise search, where k-item sets are used to explore (k + 1)-item sets.

scan database D to form \( L_1 = \{ \text{frequent 1-itemsets} \} \);

\( k = 2; \) \( k \) is the length of the item sets \$/

While \( L_{k-1} \neq 0 \) do begin \$/ association generation \$/

for each pair of \( l_{k-1}, l_{k-1} \in L_{k-1} \) and \( l_{k-1} = l_{k-1} \)

where

their first \( k - 2 \) items are the same do begin

construct candidate item set \( c_k \) such that its first \( k - 2 \) items are the same as \( l_{k-1} \), and the last two items are the last item of \( l_{k-1} \) and the last item of \( l_{k-1} \);
if there is a length k −1 subset sk−1 ∈ Ck and sk−1 ∉ /Lk−1 then remove ck; /* the prune step */
else
add ck to Ck;
end

scan D and count the support of each ck ∈ Ck;
Lk = {ck|support(ck) ≥ minimum_support};
k = k +1;
end
for all lk,k > 2 do begin /* rule generation */

for all subset am ∈ lk do begin
conf = support(lk)/support(am);
if conf ≥ minimum_confidence then begin
output rule am →(lk − am),
with confidence = conf and support = support(lk);
end
end

3.2 CLASSIFICATION:
Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of training set of data containing observations (or instances) whose category membership is known. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term “classifier” sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. Classification can be thought of as two separate problems ─ Binary and Multiclass classification. In binary classification, a better understood task, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes [13]. Since many classification methods have been developed specifically for binary classification, multiclass classification often requires the combined use of multiple binary classifiers. This is supervised learning. The class will be predetermined in training phase.

3.3 CLUSTERING:
Clustering which maps data items into groups according to similarity or distance between them. There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired. In general, clustering methods may be divided into two categories based on the cluster structure which they produce [14].

3.3.1 Non Hierarchical
The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap. These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive, and the less common agglomerating method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar however the threshold of similarity has to be defined.

3.3.2 Hierarchical – Connection Oriented
The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common divisive methods begin with all objects in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster. Some of the important Data Clustering Methods are described below.

4. CONCLUSION
Data mining methods are capable of extracting patterns automatically and adaptively from a large amount of data. Various methods related to intrusion detection system are studied briefly. This survey paper states the methods and techniques of data mining to aid the process of Intrusion Detection. The concept intercepting these two different fields, gives more scope for the research community to work in his area. New approaches will enhance existing interference detecting system and it will be a stepping stone to built effective IDS as the networks are growing rapidly in this global village.

REFERENCE:


