Use of Data Mining Techniques in Intrusion Detection – A Survey

Jasmeen Kaur Chahal\textsuperscript{1} & Amanjot Kaur\textsuperscript{2}
\textsuperscript{1}Mtech (Student), BBSBEC, FGS, India
\textsuperscript{2}Assistant Professor, BBSBEC, FGS, India

Abstract: Regardless of developing data framework broadly, security has stayed one hard-hitting territory for PCs and, in addition, systems. For that another innovation can be utilized is called Intrusion Detection System. It distinguishes the intrusions like viruses, Trojan horses and so on. Data mining is utilized to clean, characterize and inspect expansive measure of system information. In this paper, the survey of use of different data mining techniques for the detection of intrusions in the network has been done.

Keywords: security, innovation, intrusion detection system, intrusions, data mining, viruses, trojan horses.

Introduction

PC based data Systems assumes a central part in of everybody’s life. To connect our PC to the Internet, we build the danger that somebody might introduce malicious projects and utilize it to assault different machines on the Internet by handling it remotely. To keep these intrusions, for example, firewalls, client verification, data insurance and information encryption have neglected to totally shield systems and systems from the developing and refined assaults and malware. In this way, to shield the PC from different assaults and infections Intrusion Detection Systems (IDS) are composed. An IDS is a gadget which examines all inbound and outbound system action and delivers reports to the administration station. A secured system must give the accompanying:

1. Data Confidentiality: Data that is being exchanged through the system ought to be accessible just to those that have been legitimately authorized.
2. Data Integrity: Data ought to keep up their honesty from the minute they are transmitted to the minute they are really gotten. There should be no data loss from random events or malicious activity.
3. Data Availability: The network should be tough to Denial of Service attacks.

Intrusion Detection System

It is a gadget or a automated system that screens system or system activities for malicious activities or any abnormal activity which is against the policy and produces reports to an administration station. It is a blend of equipment and programming i.e. hardware and software. The use of IDS in network has been shown in Fig.1. In the figure, there are two IDS systems; one inside the network and the other outside.

The IDS devise continuously check the systems in the network and monitor servers and send the report of any abnormal activity to the network administration. There are two types of IDSes:

- Network based: Network Intrusion Detection Systems are put at a vital point or indicates inside of the network monitor activity to and from all gadgets on the network. It performs an examination of passing movement on the whole subnet, and matches the activity that is gone on the subnets to the library of known attacks.
- Host based: Host-based Intrusion detection system keeps running on individual hosts or gadgets on the network. It takes a preview of existing framework records and matches it to the past depiction. In the event that the basic system files were adjusted or erased, an alert is sent to the head to investigate.

![Figure 1. Intrusion Detection System in Network.](image-url)
Types of Intrusion Detection System

Anomaly / Statistical Detection: In anomaly based detection, the behavior of normal data has been stored in the library. If there is any activity which is not occurred previously, then the report of that activity is sent to the network administration. Anomaly based IDS can detect the attacks which are previously unknown through statistical analysis. It is also called behavior based detection technique as it detects the normal and abnormal behavior of user.

Misuse / Signature Detection: In misuse detection approach, it saves the patterns of the abnormal activity. The pattern of these abnormal activities is called signatures. Therefore, it is also called misuse-based detection. The drawback of this technique is it detects the known attacks only. Advantage of this technique is it supports the fast detection and low rate of false alarm.

Types of attacks

Probing: Probing is such type of attack, in which system is under the scanning of the intruder and checks the weaknesses and liability of the system, so that later the intruder can exploit or affect the system.

DoS Attack: In this type of attack, the attacker tries to make the resources and memory too busy, so that the legitimate user cannot access the system. This attack slows down the system or shut down the system and therefore, the corresponding authorized user is unable to access the system or service.

User to Root Attack (U2R): In this kind of attack, the attacker is used to access the system as pretending to be a normal user and then gain the privileges of the main user and affect the liability of system.

Eavesdropping Attack: In this attack, the attacker capturing the private data from the network, and read the data and gets the private information like passwords, confidential information etc. This type of attack is called network layer attack.

Remote to User Attack (R2U): In this kind of attack, the remote user can send the data to the system or access the system through internet, on which that remote user doesn’t have access to the system and affect the system privileges which a local user have on the system.

Man-in-the-Middle Attack: In this kind of attack, the attacker secretly gets the data which has been transmitted between two parties secretly. The attacker can alters the data which is sent from sender to receiver and the two original parties have no idea about this data corruption. So, the two parties think that they are communicate to the targeted party but actually the whole conversation is controlled by the attacker.

Current IDS have various considerable drawbacks namely false positive and false negative. False positive is when there is no attack occurred but the alert has been raised. False negative is when there is no alert raising when an attack has been occurred.

Data Mining Technology

Data Mining Technology has risen as a means for distinguishing patterns and trends from huge quantities of data. It is a withdrawal of concealed predictive data or learning from vast databases. Fig.2 defines the process of data mining. The first step is the gathering of data. After that the mining operation has been applied on the data and gets the result. The whole procedure of data mining is the repeated execution of these three steps. As the assortment of data is very complex because the information comes from different processes or services like log information, alarm massages etc. network activity is tremendous, so the information analysis is hard. The data mining technology has the capability of extracting large databases; it is of great importance to use data mining techniques in intrusion detection [7].

![Data Mining Diagram](Image)

**Figure 2. Process of Data Mining.**

The main advantage of data mining techniques in intrusion detection system is it detects the normal and abnormal data from vast raw data. Different data mining techniques are used to get the accurate results. These data mining techniques helps in get the comparison between normal and anomaly data by acquiring a model. Also, this is not necessary that different data mining technique is applied on different data, instead same data mining technique is used to process the different data. Various data mining techniques has been used by different researchers to get the accurate results. These techniques are: classification, clustering, association rule mining and so on.

A. Classification

A Classification is technique of predicting a value into categories i.e. classes. It takes the small data and
assigns it into a particular class. It extracts models defining important data classes. Such models are called classifiers [9]. Classification is based on predicting an outcome based on input. A corresponding algorithm of classification processes the training data and gets the results or goals. This goal is also called prediction attribute. The main aim of algorithm is to build the relationship between the input results and make it possible to predict the results.

The prediction set, which is not seen yet, which contains the same set of attributes, the algorithm takes the inputs and produces the results. The accuracy rate of prediction defines that corresponding algorithm is good or not [10]. The comparison of different classification techniques has been done by the authors of [3]. Classification is used for misuse detection and anomaly detection, but it is more commonly used for misuse detection.

B. Clustering

Clustering is a technique of making classes of the instances of same features. These classes are called the clusters. The features of instances in one cluster are different from the instances of other cluster. As classification strategy is less efficient in the field of intrusion detection. Since the measure of accessible network data is too vast, human naming is tedious and expensive. Henceforth clustering methods can be useful for classifying network data for identifying intrusions.

C. Association Rule Mining

Association Rule Mining is a guideline which implies certain relationships associate among a set of objects in the database. An association rule has two parts, an antecedent and a consequent. An antecedent is an item found in the data. A consequence is an item that is found in a mix with the antecedent. Association rules are thus made by examining the data for continuous patterns and using the criteria support and confidence to distinguish the most critical relationships. Apriori was the first scalable algorithm produced for association rule mining. The authors of [2] use the Apriori algorithm to find frequent item set.

The comparison of four algorithms has been done. These are:

- Naïve Bayes
- Hoeffding Tree
- Accuracy weighted ensemble
- Accuracy ensemble

The experimental results show that accuracy weighted ensembles gives higher accuracy than other classifiers but takes little bit more time whereas hoeffding tree gives less accuracy in less time.

Related Work

[1] created a cloud intrusion detection system to find the masquerader attacks. Masquerader is an attacker who masquerades a lawful user in the wake of abusing the user account. In that case, the firewalls or validation protocols are useless because, subsequent to logging as a legitimate user, an attacker can misuse any user benefit. The authors of this paper used association rule mining technique of data mining. Apriori-TID algorithm has been used because it is suitable as data mining is used for the transactional database. CIDD dataset has been used to evaluate the results. Also, the comparison has been done between different datasets: DARPA (1998), KDD (1999), SSENet (2011), CIDD (2012).

[2] Designs an intrusion detection system which is divided into two parts. It contains two data mining techniques: association rules and clustering. The first step is association phase in which frequent item set are produced by Apriori algorithm. The second step is a clustering phase in which clusters are created by K-Means. This technique uses the standard KDD99 intrusion detection contest dataset. The authors define that this IDS technique can detect the attacks/intrusions and classify them into different categories: U2R (User to Root), probe, R2L (Remote to Local) and Daniel of Service (DoS). The execution time is 120 ms (approx) that an approach takes to produce results. The CPU uses is 74% (approx) and memory usage is 54% (approx).

[3] used the classification techniques to detect intrusions. The authors compare the four classification algorithms and apply it to NSL-KDD datasets and compare their results. The proposed system is only intended to improve intrusion detection efficiency, not to prevent intruders.

The comparison of four algorithms has been done. These are:

- Naïve Bayes
- Hoeffding Tree
- Accuracy weighted ensemble
- Accuracy ensemble

The experimental results show that accuracy weighted ensembles gives higher accuracy than other classifiers but takes little bit more time whereas hoeffding tree gives less accuracy in less time.
[4] define the IDS which contains the clustering technique of data mining. This paper proposes using machine learning and the K-means data mining algorithm to develop an IDS model with higher efficiency rate and lower false alarms. Results of K-means clustering showed that a higher efficiency rate is achieved when the correct number of clusters is applied and increasing or decreasing the cluster beyond the number of data types only lessons the efficiency of the model.

[5] define NBA (Network Behaviour Analysis) approach for intrusion detection system. The authors applying data mining techniques on network traffic data is a solution that helps develop better intrusion detection systems. Authors suggested that combination of both approaches may overcome the limitations in current IDS and leads to high-performance ones. [6], this paper describes a novel fuzzy class-association-rule mining method based on genetic network programming (GNP) for detecting network intrusions. By combining fuzzy set theory with GNP, the proposed method can deal with the mixed database that contains both discrete and continuous attributes and also extract many important class association rules that contribute to enhancing detection ability. Therefore, the proposed method can be flexibly applied to both misuse and anomaly detection in network intrusion detection problems.

Experimental results with KDD99Cup and DARPA98 databases from MIT Lincoln Laboratory show that the proposed method provides competitively high detection rates compared with other machine-learning techniques and GNP with crisp data mining.

**Comparison**

As in Literature Survey, researchers use different algorithms to detect intrusions. The comparison of different algorithms is elaborated in Table 1.

| Table 1. Comparison of different data mining techniques for intrusion detection. |
|----------------|--------------------------------|-------------|----------------|
| Techniques     | Description                                           | Algo.       | Adv.                        |
| Classification | Classify all the network movement into either normal or malicious. It takes every instance of a dataset and assigns it to a specific class. | Support Vector Machine (SVM), Genetic Algorithm, K Nearest Neighbor, Neural Network, Decision Tree, Fuzzy Logic etc. | 1) Useful for both misuse and anomaly detection.  
2) Also used in artifice detection, retailing, analytical modeling, manufacturing, and medical analysis. | 1) Less efficient in the field of intrusion detection. |
| Clustering     | Clustering is the process of naming data and assigning into groups i.e. clustering is a division of data into groups of similar objects. | K-means clustering, hierarchical clustering, distribution based clustering, density based clustering etc. | 1) Better than classification technique [7].  
2) It can discover complex intrusions over a different time period. | 1) Clustering includes dependence on initial centroids, dependence on number of clusters and degeneracy. |
| Association Rule Mining | Association rules are made by breaking down the data for successive patterns and using the criteria support and confidence to distinguish the most vital relationships. | Apriori, Apriori TID, Apriori Hybrid, FP-Growth etc. | 1) Used for large datasets.  
2) Used where switching consumes time. | 1) Obtaining non-interesting rules.  
2) A Huge number of discovered rules.  
3) Low algorithm performance. |
Conclusion

Application of Data Mining Technology in Intrusion Detection System (IDS) is an emerging trend. In this paper, we describe different data mining techniques used for detecting intrusions. The IDS is combined with the data mining techniques and algorithms detect the threats and give an immediate response to the user. Misuse detection techniques are not sufficient for identifying unknown attacks. For detecting unknown attacks, we need to go for anomaly detection. After the survey of different papers, we conclude that the most commonly used data mining techniques are Classification, Clustering, and Association Rules.

References

[9] J. Han, M. Kambar and J. pe, Data Mining Concepts and Techniques.
[10] F. Voznika and L. Viana, "Data Mining Classification".