ISSN No. 0976-5697

Volume 3, No. 1, Jan-Feb 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Data Mining, Intrusion Detection System – A Study

Kandukuri Chandrasena Chary
Associate Professor,
Department of MCA, Sree Chaitanya Institute of Management & Computer Sciences
Karimnagar, Andhra Pradesh, India
chandrasena.mtech@yahoo.com

Abstract: in today's world network is the business driven by business needs, demands notified by the enterprises. With the pre-established network paths the enterprises do the business transactions. These pre-established network paths, information access is possible and this leads less protection of data from un-trusted users. To protect the data against un-trusted users, in the network a new system is introduced and it is part of every enterprise i.e. Intrusion Detection System. This system is used to identify the intrusions. Data mining playing a key role in strengthen the Intrusion Detection System and proven that it is the best way to detect the intrusions. This paper reviews the intrusion detection system and the mining concepts like association, classification and clustering methods in finding intrusions.

Keywords: Intrusion Detection System, IDS, Data Mining, intrusion, Denial of Service.

I. INTRODUCTION

Intrusion Detection System playing a crucial role in detecting intrusions from all parts of the network. Now a day's most of the organizations are business oriented and the business spreads across the globe. To do the business transaction they depend on the network applications via internet. This provides the intruders to access the resources of the organizations and information. From the organization point of you these are becoming vulnerable to a wide variety of cyber threats as the cost of the information processing and internet accessibility increases and the rate of cyber attacks is doubling every year in recent times. Therefore it is necessary to protect the information from such kind of attacks. To protect the business transactions and information the Intrusion Detection System (IDS) is introduced in the network to detect novel network attacks. But, according to the survey of IDS [1], the existing IDS are not providing the results at reliable level due to inability of detecting new or altered attacks.

The Intrusion Detection System is the process of monitoring and detecting the events or attacks occurred in a computer system or network and analyse the attacks for possibility of effectiveness of the system or network. The effective areas are Data Integrity, Data Confidentiality and Denial of Service. The need for effective intrusion detection mechanisms as part of a security mechanism for computer systems was recommended by Denning and Neumann [1].

The major functionality of IDS is to preserve data integrity and system availability from attacks and it is a combination of software and hardware that attempt to perform intrusion detection. The process of IDS is to collect intrusive related data from network and analyse it for intrusion. The components of Intrusion Detection System are

a. ***Signature based Detection*** – This type of system monitors packets in the Network and compares with pre-configured and pre-determined attack patterns known as signatures. The signature based detection work well for known threats not for unknown threats. The issue is that there will be lag between the new threat discovered and Signature being applied in IDS for detecting the threat. During this lag time your IDS will be unable to identify the unknown threat.

b. ***Anomaly based Detection*** – An Anomaly is a system for detecting computer intrusions and misuse by monitoring system activity and classifying it as either *normal* or *anomalous*. The classification is based on heuristics or rules, rather than patterns or signatures, and will detect any type of misuse that falls out of normal system operation. This is as opposed to signature based systems which can only detect attacks for which a signature has previously been created.

c. ***Denial of Service (DoS) Detection*** - is an attempt to make a computer resource unavailable to its intended users. The objective of DoS and Distributed DoS attacks is to deny legitimate users access to critical network services. This system compares current traffic behaviour with acceptable normal behaviour to detect DoS attacks where normal traffic is characterized by a set of pre-programmed thresholds. This can lead to false alarms or attacks being missed because the attack traffic is below the configured threshold [1].

The goal of IDS is to provide high rates of attack detection with very small rates of false alarms. Basically the alarms classified into two categories.

i. False Positive – occurs when IDS misinterprets normal attack as abnormal attack.

ii. False Negative – occurs when IDS misinterprets abnormal attack as normal attack.

The following diagram describes IDS architecture

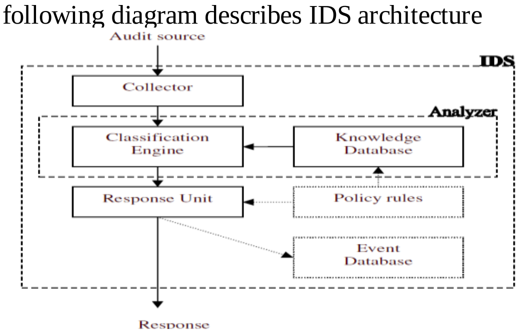

Figure 1: An IDS Model [11]

A. *Basic IDS Terminology:*

a. Console – The user interface of IDS
b. Denial of Service – against network and system overload
c. Event – The internal message of an IDS
d. False negative
e. False positive
f. Kernel – Centre of operating system
g. Manager – Collect events from sensors
h. HIDS – Host Based IDS
i. IPS – Intrusion Prevention System
j. NIDS – Network Based IDS
k. Sensor – Information gathering unit
l. Signature – formula that describes attack
m. Snort: Open source network-based IDS [11].

This diagram describes the location of IDS in the network.

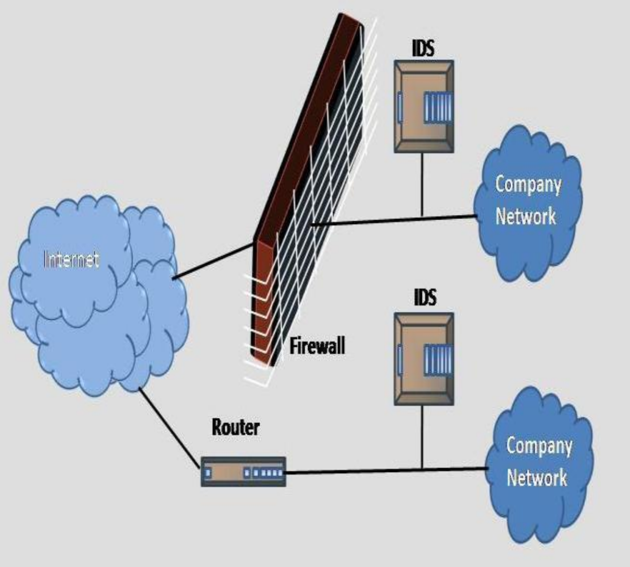

Figure 2: IDS Location.

B. *Drawbacks of Intrusion Detection System:*

Intrusion Detection Systems (IDS) have become a standard component in security infrastructure as they allow network administrators to detect policy violations. These policy violations range from external attackers trying to gain unauthorized access to insiders abusing their access. Current IDS have a number of significant drawbacks:

a. Current IDS are usually tuned to detect known service level network attacks. This leaves them vulnerable to original and novel malicious attacks.

b. ***Data Overload***: This is another aspect which does not relate directly to misuse detection but is extremely important is how much data an analyst can efficiently analyze. That amount of data he needs to look at seems to be growing rapidly. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of record per day.

c. ***False Positives****:* A common complaint is the amount of false positives IDS will generate. A false positive occurs when normal attacks is mistakenly classified as malicious and treated accordingly.

d. ***False Negatives***: This is the case where an IDS does not generate an alert when an intrusion is actually taking place

C. *Solution:*

Data mining techniques improves the intrusion detection system by addressing the above mentioned problems

D. *The Role of Data Mining:*

Data Mining is the process of extracting useful knowledge from large volumes of data integrated from various data sources. It is one of the hot topics in the field of knowledge extraction from database. Here are a few specific things that data mining might contribute to an intrusion detection project.

a. Remove normal activity from alarm data to allow analysts to focus on real attacks.
b. Identify false alarm generator and "bad" sensors signatures.
c. Find anomalous activity that uncovers real attack
d. Identify long, ongoing patterns (different IP address, same activity)

To accomplish these tasks, data miners employ one or more of the following techniques:

a. Data summarization with statistics, including finding outliers.
b. Visualization: presenting a graphical summary of the data.
c. Clustering of the data into natural categories.
d. Association rule discovery: defining normal activity and enabling the discovery of anomalies.
e. Classification: predicting the category to which a particular record belongs.

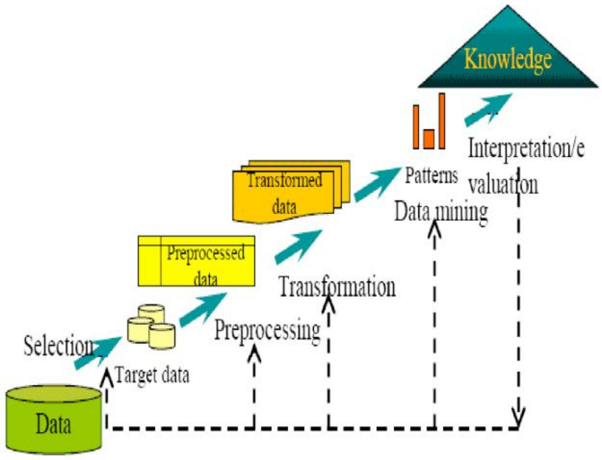

Figure 3: Data Mining Approach

II. IDS MODELS USING DATA MINING

The models of IDS based on data mining mainly includes Data collection, Data Mining, Pattern matching and Decision making [3]. According to this, the mining engine process network data into data records. Next, the mining algorithms analyze data records to know normal or abnormal rules in the record. If new rules are found, generate and store the rules. In the next step comparing the rules with newly generated rules then analyze the rules by intelligence decision making module to decide whether intrusion has taken place or not.

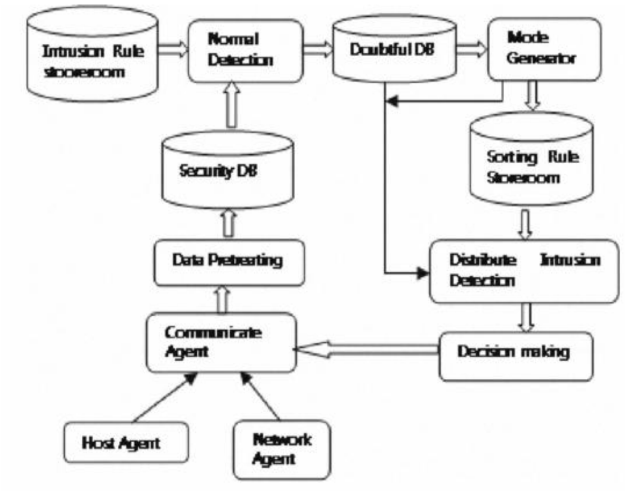

Figure 4: IDS Model based on Data Mining [3]

With the available working procedures of data mining in IDS is first, collect monitored data and covert to formatted data suitable to mining process. Second, with the available data build clustering or classification model. This model assists to identify the intrusion during the intrusion detection stage.

Recently Data Mining is becoming important element in IDS by using the approaches of data mining like clustering, classification, association and outlier analysis. These approaches are used to gather intrusion related data

III. INTRUSION DETECTION USING ASSOCIATION RULE

Association Rule Mining is one of the most important, frequently used and fundamental technique in data mining which discovers the association and correlation among the item sets in large database. Apart from this it is specifically designed for use in data analysis. This rule provides the information in the form of “if-then” statement. The association rule available either in Apriori model or frequent pattern growth model [5]. The format of association rule is “X=>Y” where X & Y are the itemsets can be interpreted as presence of X itemset in a particular transaction implies the presence of Y itemset. The extraction of such association from transaction data base is referred to as association rule mining. The association rule mining is twostep processes [6], in the first step identify the frequent itemsets while in the second step extract the association rule from set of frequent itemsets.

In the process of applying association rule mining in IDS, first collect audit data from network. This data is processed into data records in a text file. Each entry in the text file corresponds to a connection entry associated with source IP and target IP address. Next, the audit data is processed for data mining and is split into two files, the training set and test set [8].

The basic steps for incorporating association rule for intrusion detection as follows. (1) First network data need to be formatted into database table where each row is an audit record and each column is a field of the audit records. (2) Maintain correlation among network data and consist behaviours can be captured in association rules. (3) Rules can be continuously merged from a new run to the aggregate rule set of all previous runs [5].

The motivation for applying the association rules algorithm to audit data are (1)Audit data can be formatted into a database table where each row is an audit record and each column is a field (system feature) of the audit records (2). There is evidence that program executions and user activities exhibit frequent correlations among system features (3). We can continuously merge the rules from a new run to the aggregate rule set (of all previous run) [12].

IV. INTRUSION DETECTION USING CLUSTERING

Clustering is one of the data mining methods used in IDS to gather intrusion related data. Clustering is the process of labelling data and assigning it into groups. Clustering can group new data instances into similar groups. These groups can be used to increase the performance of existing classifiers. Clustering can be similarity-based and centroid-based for assigning labels to data.

Clustering is an unsupervised machine learning mechanism for finding patterns in unlabeled data with many dimensions. K-means clustering is used to find natural groupings of similar alarm records. Clustering discovers complex intrusions occurred over extended periods of time and different spaces, correlating independent network events. The basic steps involved in identifying intrusion **-** (1) Find the largest cluster; (2) Sort the remaining clusters in an ascending order of their distances to the largest cluster; (3) Select the first K1 clusters; (4) Label all the other clusters as attacks [5]. After clustering, heuristics are used to automatically label each cluster as either normal or attacks.

In [9] authors explore the use of clustering in the field of Intrusion Detection System. The focused clustering techniques are; (1) Optimized Sampling with Clustering Approach for Large Intrusion Detection Data; (2) Clustering Based Method for Unsupervised Intrusion Detection; (3) Clustering Algorithm to Enhance the Performance of the Network Intrusion Detection System. To remove the problems in existing techniques, a new model is proposed by [9] which are based on feature selection as a firs phase, K-Mean clustering as a second phase, classification as third phase, and final phase is evaluating the performance in terms of precision and recall.

In paper [10] authors proposed a clustering model for intrusion detection for identifying anomalous events based on simple K-Means for analysing network flow data using any flow data attribute such as IP address, port, protocols to detect anomalies. In this approach a Netflow collector for collecting net flow records from the network traffic. This data set is pre-processed and then filtered based on key parameters like IP addresses, ports and protocols. Data mining technique is applied over the filtered data.

V. INTRUSION DETECTION USING CLASSIFICATION

Classification is similar to clustering in that it also partitions network audit records into distinct segments called classes. In the classification process the end user must know how classes are defined. Classification categories the data records in a predetermined set of classes used as attribute to label each record; distinguishing elements belonging to the normal or abnormal class. Classification approach can be useful for both misuse detection and anomaly detection, but it is more commonly used for misuse detection. Classification steps of intrusion detection are proposed in [5]

VI. CONCLUSIONS

The Intrusion Detection System is a key part of network to detect intrusions during flow of data from network to network. With IDS, the system is available with intrusion free by using the techniques like signature based or anomaly based detection methods. The performance of IDS can be improved by using Data mining techniques like clustering, association and classification when network audit data is overloaded. Data mining helps in gathering intrusion related data which is used for further decision making. In future work, by using artificial intelligence, neural network and focusing on more number of attributes network audit data we can increase the efficiency of Intrusion Detection System.

VII. REFERENCES

[1] Naveen NC, Dr R. Srinivasam, Dr S.Natarajan, “Research Directions in Intrusion Detection, Prevention and Response System – A Survey”, IFRSA International Journal of Data Warehousing & Mining |vol1|issue 1|Aug 2011, 95-100.

[2] Lie Wenjun, “An Security Model: Data Mining and Intrusion Detection”, IEEE 2nd International Conference on Industrial and Information System 2010, 448-450.

[3] Umesh Sehgal, ”DATA MINING: An Overview”. International Journal of Data Warehousing & Mining|Vol1|issue 1|Aug 2011, 69-74.

[4] Chang-Tien Lu, Arnold P.Boedihardjo, Prajwal Manalwar, “Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems”, 2005 IEEE, 512-517.

[5] Kamal K Sethi, D K Mishra, Gopal Solanki, Bharat Mishara, “Key Issues of Security and Integrity in Third Party Association Rule Mining”, 2009 IEEE Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09, 337-340.

[6] S. Sathya Bama, “Network Intrusion Detection using Clustering: A Data Mining Approch”, International Journal of Computer Applications |Vol 30-No 4, Sep’ 2011, 14-17.

[7] Flora S. Tsai, “Network Intrusion Detection Using Association Rules” 2009 ACADEMY PUBLISHER, International Journal of Recent Trends in Egineering, Vol2, No.2, Nov’ 2009, 202-204.

[8] Kusum Kumari Bharti, “Intrusion Detection using Clustering”, International conference ACCTA-2010, IJCCT Vol.1 Issue 2,3,4; 2010, 158-165.

[9] Mayank Pal Singh, Subramanian N, “Visualization of Flow Data Based on Clustering Technique for Identifying Network Anomalies”, 2009 IEEE Symposium on Industrial Electronics and Applications, October 4-6, Kuala Lumpur, Malaysia, 973-978.

[10] Intrusion Detection System – Technologies, Weaknesses and Trends by Martin Arvidson, Markus Carlbark, Stockholm 2003.

[11] Wenke Lee and Salvatore J. Stolfo, Columbia University, “Data Mining Approaches for Intrusion Detection”, Proceeding of 7th USENIX Security Symposium San Antonio, Texas, January 26-29, 1998.