ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

Literature Survey on Data Mining Techniques For Intrusion Detection

[I]Arpitha J, [II]Nagaraj Naik S
[I]M.Tech Student, [II]Senior Assistant Professor
[I,II]Dept.of CSE, Mangalore Institute of Technology and Engg., Mangalore, Karnataka, India

Abstract
Intrusion detection is the act of detecting activities in the network that compromise the confidentiality, availability, integrity of a resource. Data mining is one of the technologies applied to ID to invent a new pattern from the massive network data as well as to reduce the strain of the manual compilations of the intrusion and normal behavior patterns. This paper presents various techniques used to detect intrusions along with their advantages and disadvantages.

Keywords
Data Mining; Intrusion Detection; Intrusion; Types of Intrusions; Clustering.

I. Introduction

The internet has become a part of daily life and an essential tool today. As more and more sensitive data continues to get stored and manipulated online; the need for an increase in security of network systems is getting more and more importance day by day. The following 3 functionalities must be provided essentially by any secure network [1].

- ***Data confidentiality:*** Data access must strictly be done by authorized users. Eavesdroppers and intruders must mot gain any important information.
- ***Data availability:*** Authorized system users must be able to access and use any resource at any point in time.
- ***Data integrity:*** Corruption and data loss of information must be prevented. Exactness of data must be preserved.

II. Intrusion and its Types

Intrusion is an illegal act of entering, seizing, or taking possession of another's property without any permission. It means a code that disables the proper flowing of traffic on the network and steals the information from the traffic. Some of common names for the intrusions are viruses, Trojan horse etc. Some of categories of intrusions are described below [5]:

DoS Attack*:* **Denial-of-service** *attack is* a type of attack on a network that is designed to bring the network to its knees by flooding it with useless traffic. Usual target for such kinds of attacks are high profile web servers such as banks. Though DoS attacks do not typically result in the theft or loss of information they can cost the victim a great deal of time and money.

Remote to User (R2L): Here an attacker tries to gain access to the local machine from a remote machine by some unauthorized means. Social engineering is one such attack.

Probes**:** It is a class of attacks where an attacker continuously scrutinizes a network until he finds all the vulnerabilities present. Attacks are then staged by exploiting these loopholes.

User to Root (U2R): Here an attacker tries to attack where a local user on a machine is able to obtain privileges normally reserved for the UNIX super user or the Windows NT administrator.

III. Intrusion Detectin System

An Intrusion Detection System (IDS) is a defense system which inspects the activities in a system for suspicious behavior or patterns that may indicate system attack or misuse and then notify intrusion prevention system (IPS) or network security administrator so that suitable actions can be taken against the attacks. Following are the 2 important approaches to detect intrusions [4].

A. Misuse detection

In Misuse detection patterns for different malicious behaviours are built first, and then the attacks are detected based on these predefined patterns. Misuse detection is very effective in avoiding an immense amount of false alarms and provides a great accuracy. However, misuse detectors can only detect attacks whose signatures are known. Any variations of the common attacks go undetected. Signatures of new attacks must be constantly updated.

B. Anomaly detection

In anomaly detection, a normal profile which describes the behaviour of the system under normal conditions is constructed in advance. Any significant aberrations from such expected behaviour are reported as possible attacks. The major advantage of this approach is that with fewer details unusual behavior can be easily detected, thereby effectively reducing the storage and maintenance cost. But it requires a large amount of "training sets "to effectively characterize normal behavior. Another shortcoming of anomaly detection is its high false alarm rate.

IV. Categories of IDS

Intrusion detection is a combination of both software and hardware used to detect intruder activity by using a set of techniques and methods at the network and host level.

- ***Host-based IDSs (HIDS):*** resides on the system it monitors and tracks changes made to files and directories. It takes a snap shot of existing system files and matches it to the previous snap shot. If the critical system files were modified or deleted, the alert is sent to the administrator to investigate.
- ***Network-based IDSs (NIDS):*** examines data exchanged between computers. Raw network packets that travel between devices connected in a network serve as a data source. The IDS typically uses a network adapter in promiscuous mode that listens and analyses all traffic in real-time as it travels across the network.

V. Drawbacks of Current IDS

Presently used IDS suffer from many drawbacks. The prime among them are [1]:

- ***Unknown attacks:*** Traditional signature based method

requires an extensive knowledge of signature of previously known attacks. This method matches the monitored events with the signatures stored in the database to detect intrusion. SNORT is one such example. Even though accuracy is greater this method is vulnerable to unknown and novel attacks.

- ***Data overload:*** Handling huge data daily and effectively analyzing them can be cumbersome.
- ***False positives:*** In this scenario a normal data is mistaken as a malicious one and suitable protection mechanisms are enforced against it.
- ***False negatives:*** In this case, no alert is generated to detect an intrusion misinterpreting it to be a normal one.

VI. Data Mining Techniques to Detect Intrusion

Data mining is the analysis of a large amount of data for relationships and useful patterns that have not previously been discovered. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists).

C. Classification

In Classification [3], every single data of a data set is allotted to a particular class. Data classes are developed by use of models called as classifiers. Entire network traffic is either grouped under normal or intrusion classes according to their behaviour.

Classification is a supervised machine learning mechanism. It is only suitable to work with labelled data. However if the data matches a pre computed class model, training time taken is less. Example of classification based approach is decision tree and Naïve Bayes Classifier.

D. Association rule discovery

Market Basket Analysis uses this method to find associations amongst the items in the customer's cart. This helps the dealers in deciding which items must be placed together and also to which items discount should be provided so that their sales can be improved.

Let I ={I1, I2, ….., Im} be a set of items in a database D having transactions D = {t1, t2, ….., tn} where ti = {Ii1, Ii2, …., Iik} and Iij ∈ I [2].An association rule is an implication of the form $X \rightarrow Y$ where X, Y ⊂ I are the sets of items called item sets and $X \cap Y = \emptyset$[2]. The support for an association rule $X \rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$ [2]. The confidence or strength for an association rule $X \rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X[2].

Even though this approach was well suited for Market Basket Analysis it is not the best approach to detect intrusions as processing large number of rules is tiring. Also the execution time here increases with the number of attributes.

E. Machine Learning Approaches

Machine learning can be defined as the study of computer algorithms which enables the machine to improve its performance for a given set of tasks due to prior training. Machine learning techniques can change their execution approach and game plan according to newly acquired information, but the major drawback is their expensive resource requirements and complex, time consuming training requirements.

Bayesian Approach, Neural Networks, Fuzzy Logic, Genetic Algorithms and Support vector machines are some of the machine learning techniques [2].

F. Clustering

Clustering is the process of assigning the data into groups based on similarity. Each group is called as a cluster. This process ensures that intra-cluster distance is less and inter-cluster distance is more. But the trait of clustering method to force the data into one or more clusters makes it less favorable [3].

K-Means [4] is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a given set of *n* data objects in *k* clusters, where *k* is the number of desired clusters and it is required in advance. K-Mediod clustering algorithm overcomes shortcomings of K-Means: that is number of clusters dependency, dependence on initial centroids and degeneracy. This is achieved by using a Mediod instead of a cluster.

G. Hybrid Learning Approaches

Variety of methods such as fusion of clustering and classification techniques can be use to formed a hybrid learning approaches. This technique offers high detection rate and low alarm rate [3].

Most common example is a combination of Naïve Bayes Classifier and K-Means [6]. Here after grouping the data into suitable clusters, classifier is applied for classification purpose

VII. Conclusion

Research on Intrusion Detection has been going on since the 1980s. With the emergence of newer and more in penetrable attacks providing security to our systems has become our upmost concern. There is a need for an effective system which detects malicious attacks with ease and accuracy avoiding false alarms. This paper describes intrusions and their types along with the limitations of current intrusion detection systems. This paper also illustrates how data mining aids in intrusion detection process and lists various techniques applied and evaluated by researchers.

VIII. Acknowledgment

I am very thankful to my guide Mr. Nagaraj Naik, Sr. Assistant Professor, Department of Computer Science and Engineering, Mite for his cordial support, valuable information and guidance, to prepare this paper and also thankful to Prof. Dr. Nagesh H R, Head of the Department, Computer Science and Engineering, for his valuable and constructive suggestions during the planning and development of this work.

References

[1] Reema Patel, Amit Thakkar, Amit Ganatra," A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[2] Manasi Gyanchandani, J.L.Rana, R.N.Yadav," Taxonomy of Anomaly Based Intrusion Detection System: A Review", International Journal of Scientific and Research Publications ISSN 2250-3153, Volume 2, Issue 12, December 2012.

[3] RavindraThool, Kapil Wankhade ,Sadia Patka," An Overview of Intrusion Detection Based on Data Mining Techniques", International Conference on Communication Systems and Network Technologies, 2013.

[4] Poonam Dabas, Rashmi Chaudhary, "Survey of Network Intrusion Detection Using K-Mean Algorithm", International

Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[5] Amanpreet Chauhan, Gaurav Mishra, Gulshan Kumar," Survey on Data Mining Techniques in Intrusion Detection", International Journal of Scientific & Engineering Research Volume 2, Issue 7, July-2011.

[6] Z. Muda, W. Yassin, M.N. Sulaiman, N. I Udzir,"A K-Means and Naïve Bayes Approach for Better Intrusion Detection", Information Technology Journal, 648-655, 2011.

Authors Profile

Miss Arpitha J completed the Bachelor's Degree in Computer Science & Engineering from Visvesvaraya technological University (VTU). Currently pursuing M.Tech degree in Computer Science & Engineering at Mangalore Institute of Technology, Mangalore under VTU, Belgaum.

Mr. Nagaraj Naik senior assistant professor MITE, Mangalore. Completed his M.Tech in computer science and engineering having 8.5 years of academic experience and his areas of interest are java programming, operating system.