A Survey on Areas in Data Mining for Intrusion Detection

Nisha P.Shetty
Department of CS&E
Canara Engineering College
Mangalore, India
e-mail: pnishashetty@gmail.com

***Abstract*—** Intrusion detection is the act of detecting activities that compromise the confidentiality, integrity and availability of a resource. Data Mining is the process of analyzing huge amounts of data to obtain useful information for the required cause. This paper presents various techniques used to detect intrusions along with their pros and cons. An efficient detection method must provide proper diagnosis of any obstruction with greater accuracy and low false alarm rate.

***Keywords*—**clustering, categories of intrusions, data mining, intrusion, types of intrusions

I. INTRODUCTION

Internet is gaining huge popularity nowadays as it aids people in various areas of their life such as business, entertainment, education etc. 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 not gain any important information.
- *Data integrity:* Corruption and loss of information must be prevented. Exactness of data must be preserved.
- *Data availability:* Authorized system users must be able to access and use any resource at any point in time.

II. INTRUSION AND ITS TYPES

Intrusion in simple terms means an illegal act of entering, seizing and manipulating other's property without any valid permissions. Some of commonly observed examples of intrusions are viruses, Trojan horse etc. Some of the categories of intrusions are described below [5]:

- *DoS Attack*: This group of attack makes a computer resource such as an Internet site or any such service from functioning efficiently temporarily or indefinitely. Usual target for such kinds of attacks are high profile web servers such as banks.
- *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.
- *User to Root (U2R):* Here an attacker having access to a normal user account on the host system abuses vulnerabilities in the system to gain root access to the system [12]. Buffer overflow error caused by regular programming mistakes [12] and faulty environmental assumptions belongs to this category.
- *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.

III. INTRUSION DETECTION SYSTEM

An Intrusion Detection System (IDS) is a defense system [16] 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 these attacks. Following are the 2 important approaches to detect intrusions [4].

A. Misuse detection

In Misuse detection patterns for different malicious behaviors 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 greater 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 behavior of the system under normal conditions is constructed in advance. Any significant aberrations from such expected behavior are treated as possible attacks to be investigated. 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 the normal behavior. Another shortcoming of anomaly detection is its high false alarm rate.

IV. Categories of IDS

Intrusion detection is a software, hardware or combination of both [15] used to detect an intruder activity by using a set of techniques and methods at the network and (or) host level [17].

A. Host-based IDSs (HIDS)

It monitors the data contained in individual computers known as hosts. In this system IDS resides on each host and is governed by the system. It detects intrusions by analyzing application logs, system calls, file-system modifications, and other host activities that related to the server computers.

B. Network-based IDSs (NIDS)

It 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 [7].

V. Drawbacks of Current IDS

Presently used IDS suffer from many drawbacks. The prime among them are [8]:

- *Unknown attacks:* Traditional signature based method requires an extensive knowledge of signatures 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 amounts of 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 Used to Detect Intrusions

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

A. 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 behavior.

Classification is a supervised machine learning mechanism [10]. It is only suitable to work with labeled data. However if the data matches a pre computed class model, training time taken is less. Examples of classification based approach are Decision Tree [11] and Naïve Bayes Classifier.

B. Association Rule Mining

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 for 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 an individual transaction ti contains {Ii1, Ii2, …., Iik} and each Iij ∈ I [2].An association rule is an implication of the form $X \rightarrow Y$ where X, Y $\subset$ 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 is 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.

C. Machine Learning Approaches

Machine learning [9] 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].

D. 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 which is required in advance [13]. K-Mediod [14] clustering algorithm overcomes shortcomings of K-Means: that is number of clusters dependency, dependence on initial centroids and degeneracy.

E. Hybrid Learning Approaches

Variety of methods such as fusion of clustering and classification techniques can be use to form a hybrid learning approach. 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.

Acknowledgment

I am very grateful to **Mr. Deepak D.,** Assistant Professor, Canara Engineering College, Mangalore for his invaluable guidance, inspiration and constructive suggestions that helped me in the preparation of this paper.

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.

[7]Brian Laing, "How To Guide-Implementing a Network Based Intrusion Detection System".

[8]Harshna , Navneet Kaur," Survey paper on Data Mining techniques of Intrusion Detection", International Journal of Science, Engineering and Technology Research (IJSETR) , Volume 2, Issue 4, April 2013.

[9]Theodoros Lappas , Konstantinos Plechrinis," Data mining techniques for (Network) Intrusion Detection Systems".

[10] Yousef Emami, Marzieh Ahmadzadeh,Mohammad Salehi,Sajad Homayoun ," Efficient Intrusion Detection using Weighted K-means Clustering and Naïve Bayes Classification", Journal of Emerging Trends in Computing and Information Sciences, Vol. 5, No. 8 August 2014.

[11] Neha Jain, Shikha Sharma, "The Role of Decision Tree Technique for Automating Intrusion Detection System ", International Journal of Computational Engineering Research (ijceronline.com), Vol. 2 Issue 4.

[12] Suresh Kashyap, Pooja Agrawal, Vikas Chandra Pandey, Suraj Prasad Keshri," Soft Computing Based Classification Technique Using KDD 99 Data Set for Intrusion Detection System ",International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 4, April 2013 .

[13] Pankaj Saxena , Vineeta Singh ,Sushma Lehri," Evolving Efficient Clustering Patterns in Liver Patient Data through Data Mining Techniques , International Journal of Computer Applications , Volume 66– No.16, March 2013.

[14] Ravi Ranjan, G. Sahoo," A New Clustering Approach For Anomaly Intrusion Detection", International Journal of Data Mining & Knowledge Management Process (IJDKP) ,Vol.4, No.2, March 2014.

[15] http://technbyte.blogspot.in/2011/11/what-is-intrusion-detection-syste-m.html.

[16] An Approach for Intrusion Detection Based Information Technology Essay:http://www.ukessays.com/essays/information-technology/an-approach-for-intrusion-detection-based-information-technology-essay.php.

[17] Syracuse University Lecture Notes for Internet Security, "Intrusion Detection System", Fall 2006.

Ms. Nisha P.Shetty has received her B. E degree in Computer Sc. & Engg. from Srinivas School of Engineering Mukka, Mangalore under VTU Belgaum. Currently she is pursuing M Tech. in Computer Sc. & Engg. at Canara Engineering College ,Mangalore. Her areas of interest are Data Mining and Database Systems.