Survey of Data Mining Approach using IDS

[1]Raman kamboj, [2]Kamal Kumar
Research Scholar, Assistant Professor
SDDIET, Department of Computer Science & Engineering, Kurukshetra Universty

***Abstract* - In our days, electronics attacks can cause a very destructive damage for nations which make necessary the use of completed security policy to minimize the potential threats. Intrusion detection is to identify attacks against a computer system. It is an key technology in production sector as well as an active area of do research. In Information Security, intrusion detection is the act of detecting actions that attempt to cooperation the privacy, integrity or availability of a resource. It plays a very significant role in attack finding, security check and network check. One of the main challenge to intrusion detection are the problem of misjudgment, misdetection and require of actual time response to the attack.**

***Keywords* - IDS ,HIDS ,NIDS, Clustring**

I. INTRODUCTION

Informally, an Intrusion Detection system is a system for raising attention towards potential misbehaviors of the system caused by external adversaries. We could think of a 'burglar alarm' in the real world as the physical analogue of an intrusion detection system in the computerized world. (Just as a burglar alarm in the real world, Intrusion Detection only deals with discovering that an intrusion might have happened into a network. A number of additional aspects related to intrusions, such as intrusion avoidance; that is, augmenting systems so to have a lower likelihood of an external attacker that successfully performs an intrusion; or intrusion tolerance; that is, augmenting systems A network intrusion attack can be any use of a network that compromises its stability or the security of information that is stored on computers connected to it. A wide range of activity falls under this definition, including attempt to de-stabilize the network as a whole, gain unauthorized so that the intended system behaviour does not change even after an intrusion; are the subject of study of different research areas.)

Intrusion Detection is a very active and important research area in the Security literature. We won't attempt to survey or categorize the research in this area, but we note that the origin of the problem is often attributed to [1] and several taxonomies and surveys can be found, for instance, in [3, 4]. Often all techniques in known intrusion detection systems are abstracted as falling under two important principles: anomaly detection, according to which traffic significantly different from normal ones can be interpreted as likely to be an attack, and signature detection, (also called misuse detection or rule-based detection), according to which traffic significantly similar to known attack traffic can be interpreted as likely to be the same attack. Both principles offer advantages and disadvantages, and many recent systems combine the two principles, rather than specifically choosing one of them. Despite the large amount of research in this area, no established common framework exists for the design and analysis of intrusion detection systems. A typical research paper in the area proceeds describing some new ideas for detecting intrusions and justifies their validity by describing a specific implementation experience where both the rate of 'false positives' and the rate of 'false negatives' are low.

II. SECURITY ISSUE IN IDS

1Network-Based IDS Network based IDS (Fig. 1.1) are best suited for alert generation of intrusion from outside the perimeter of the enterprise. The network based IDS are in sallied at various points on LAN and observe packets traffic on the Network information is assembled into packets and transmitted on LAN or Internet. N-B IDS are valuable if they placed just outside the firewalls, thereby altering personal to incoming packets that might circumvent the firewall [2]. Some Network-Based IDS take or allows taking input of "Custom signatures" taken from user's security policy, which pelmets limited detection security policy violation.

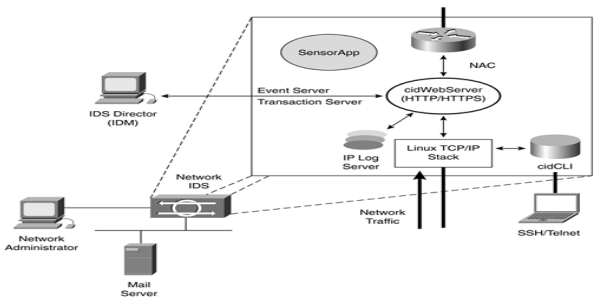

Fig 1.1 Network Based IDS

2Host Based IDS Host-based IDS (Fig. 1.2) placed monitoring "Sensors" also known as "agents" on network resources nodes to monitor audit logs that are generated by Network Operating System or application program. Audit logs contain records for events and activities taking place at individual Network resources. Because this Host-Based IDS can detect attacks that cannot be seen by Network based IDS . Such as Intrusion and misuse and misuse by trusted insider. Host-Based can overcome the problems associated with N/ W based IDS immediately after alarming the security personnel can locate the source provided by site -security policy [2].

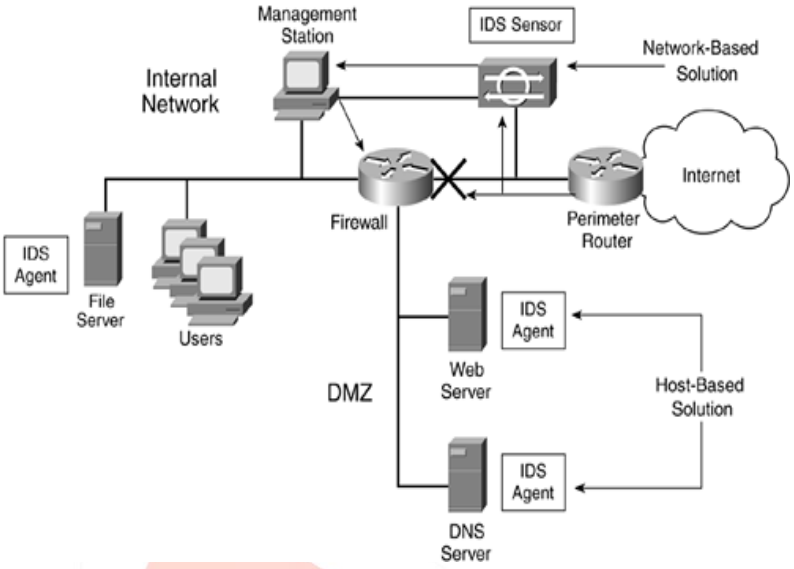

Fig 1.2 Host Based IDS

III.RELATED WORK

S.V.Shirbhate et al[2014] The study, analysis and exploration of recent development of data mining applications such as classification and clustering is one of the needs for machine learning algorithms to be applied to large scale data will lead to acquire the direction of future research. It would be future demand in IDS for detecting the intrusions in mobile network. This paper presents the comparison of different clustering techniques. Also focus on the effect of Principal Component Analysis filter on these clustered based methods.

Chakchai Soet al[2014] Due to a rapid growth of Internet, the number of network attacks has risen leading to the essentials of network intrusion detection systems (IDS) to secure the network. With heterogeneous accesses and huge traffic volumes, several pattern identification techniques have been brought into the research community. Data Mining is one of the analyses which many IDSs have adopted as an attack recognition scheme. Thus, in this paper, the classification methodology including attribute and data selections was drawn based on the well–known classification schemes, i.e., Decision Tree, Ripper Rule, Neural Networks, Naïve Bayes, *k*–Nearest–Neighbour, and Support Vector Machine, for intrusion detection analysis using both KDD CUP dataset and recent HTTP BOTNET attacks. Performance of the evaluation was measured using recent Weka tools with a standard cross–validation and confusion matrix.

Mouaad KEZIH et al [2013] Intrusions detections systems from point of view of security policy are a second line of defense; they have a supervisory role to observe the activities of our network or hosts to identify attacks in actual time. In our days, electronics attacks can cause a very destructive damage for nations which make necessary the use of completed security policy to minimize the probable threats. IDS it is a very important element to resist against this vulnerability, (KDD) CUP 99 and a Data Mining Tools Waikato Environment for Knowledge Analysis (WEKA) to combine the advantages of an intrusion detection algorithm (PART) and two techniques of Dimensionality Reduction(best first search and genetic search), to estimate our works, we applied the proposed combined technique ,and we check the results by using a several evaluations parameters.

Tayeb Kenaza et al[2010] we propose to combine a behavioral intrusion detection approach with a clustering approach in order to obtain a set of clusters with different false alerts rates. The order of these clusters with respect to their false alerts rates will be considered as an alerts prioritization. Hence, new alerts will be classified to the closest cluster and processed according to their cluster priority. Several machine learning mechanisms such as neural networks, support vector machines (SVM), decision trees, Bayesian networks, etc. have been used for the design of behavioral based intrusion detection systems (IDS).

Meng Jianliang et a[2009l Internet security has been one of the most important problems in the world. Anomaly detection is the basic method to defend new attack in Intrusion Detection .Network intrusion detection is the process of monitoring the events occurring in a computing system or network and analyzing them for signs of intrusions, defined as attempts to compromise the confidentiality. Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar, and the members from different clusters are quite different from each other. Until now, the clustering algorithms can be categorized into four main groups: partitioning algorithm, hierarchical algorithm, density-based algorithm and grid-based algorithm. Partitioning algorithms construct a partition of a database of N objects into a set of K clusters.

Z. Muda et al[2011] Intrusion Detection System (IDS) plays an effective way to achieve higher security in detecting malicious activities for a couple of years. Anomaly detection is one of intrusion detection system. Current anomaly detection is often associated with high false alarm with moderate accuracy and detection rates when it's unable to detect all types of attacks correctly. To overcome this problem, we propose an hybrid learning approach through combination of K-Means clustering and Naïve Bayes classification. With the rapid growth of network technology, a cyber crime incident has also grown accordingly. A wide range of risks and threats against uncontrolled and undefended assets such as database and web server as well as entire network system become the general concern for intruders nowadays. Gaining unauthorized access to files, network and any other serious security threat can be detected by employing Intrusion Detection System. IDS identify any activity that violates the security policy from various areas within computer and network environment.

There are two traditional IDSs used to detect intruders: signature-based detection and anomaly-based detection. A signature-based IDS match define signature with each anayzed packets on the network to detect known malicious attack as a same way like a virus scanner. These type of IDS required a frequent updating for the new signatures to keep the signature database up-to date Thus, it fails in discovering and detect an unknown attacks once the signature did not exist in its library. Unlike signature-based detection, anomaly-based detection is designed to capture any activities which are diviates the normal usage pattern called normal profile. In recent years, data mining approach have been proposed and used as detection techniques for discover unknown attacks. This approach has resulted in high accuracy and good detection rates but with moderate false alarm on novel attacks.

Abhay Kumar et al[2012] Two common data mining techniques for finding hidden patterns in data are clustering and classification analyses. Classification is supposed to be supervised learning and clustering is an unsupervised classification with no predefined classes. Clustering tries to group a set of objects and find whether there is some relationship between those objects. In this paper we have used the numerical results generated through the Probability Density Function algorithm as the basis of recommendations in favor of the K-means clustering for weather-related predictions.

IV. DATA MINING

The open source data mining framework Weka was the tool we used for testing the traditional classify, cluster and association algorithms. Weka was chosen for a variety of reasons, not the least of which is because it is "well-suited for developing new machine learning schemes." [8] It also enjoys widespread community support and has many algorithms from which to choose. In future efforts we plan to report on results generated by three types of algorithms:

- Classifying
  - UserClassifier - weak classifiers trees (a novel interactive decision tree classifier)
  - J48 Tree
- Clustering – Cobweb, SimpleKMeans
- Association – Apriori

As noted above, ARFF files are effectively textual flat files and, given the difficulties and extra preprocessing steps necessary to manipulate away the missing data, it may be advantageous to take advantage of Weka's ability to connect to a database rather than try and preprocess the flat file.

V. CLUSTER ANALYSIS METHODS IN DATA MINING

Cluster analysis is a very important data mining technology to divide the data object into several meaningful subclasses, so that the members from the same clusters are quite similar and members from different clusters are quite different from each other [7]. Therefore this method is applied for classifying log data and detecting intrusions. Clustering is an unsupervised learning technique of data mining that takes unlabeled data points and tries to group them according to their similarity. In unsupervised approach there is no need of prior knowledge about training data whereas in supervised approach, given a set of normal data need to train in order to detect whether the test data belongs to normal or anomalous behavior [6].

The general steps for clustering are: feature extraction from sample data where input is sample data and output is matrix. Then implementation of clustering algorithm to access cluster genealogy diagram i.e. to reflect all the classification. After obtaining a cluster genealogy diagram, the domain experts will decide the threshold selection according to the specific application by experience and domain knowledge. Data preprocessing, this includes standardization, integration; normalization etc. is one of the important step before applying data mining. This is also necessary precondition for normal operation of clustering [7].

Clustering algorithm can be categorized into four main groups: partitioning algorithm, hierarchical algorithm, density based algorithm and grid based algorithm as shown in fig1.

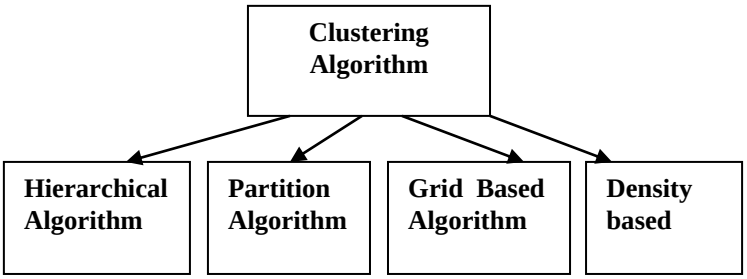

Fig4. 1. Classification of clustering algorithm

A. *Density-based methods*

Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. The general idea is to continue growing the given cluster as long as the density i.e. number of objects or data points in the neighborhood exceeds some threshold. It means that for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise or outliers and discover clusters of arbitrary shape [7].

B. Grid-based methods

Grid-based methods divide the object space into a finite
number of cells that form a grid structure. All of the clustering operations are performed on the grid structure. The main advantage of this approach is its fast processing time, which is typically dependent mainly on the number of cells in each dimension in the quantized space [1].

C. Hierarchical Cluster

Hierarchical Clustering Methods are Agglomerative hierarchical methods. This Begins with as many clusters as objects. Clusters are successively merged until only one cluster remains. Divisive hierarchical methods begin with all objects in one cluster. Groups are continually divided until there are as many clusters as objects [2].

D. *Partition Algorithm*

Partitioning algorithm divides database of N objects into K clusters. Usually start with an initial partition and then use an iterative control strategy to optimize an objective function.

VI. TYPE OF ATTACK

1. DoS: This attack can freeze the server operation and activity by acquiring all resources so that the server cannot provide any service, commonly using flooding–based schemes.
2.PROBE: This attack is used during a preparation stage for other attacks in order to gain valuable information such as enabled ports and services as well as Internet address information.
3.U2R: This attack performs a specific operation in order to penetrate into a system hole/leak such as Buffer Overflow.
4.R2L: The attack is used to take advantages of related users' safety information or configuration such as SQL Injection.
5. BOTNET: This attack is to run the computer into a bot (zombie) to perform a particular task over the Internet as administrator.

VII. CONCLUSIONS

Since the expedient data mining algorithms is available, intrusion detection based on data mining has developed rapidly. In this paper produce on intrusion detection as a process of data analysis by using the predominance of data mining in its effective use of information, this is a method that can by design produce accurate and applicable intrusion patterns from massive audit data, which makes intrusion detecting system, can be applied to any computer background. This approach has become a accepted topic of research, in the field of inter discipline of network security and artificial intelligence .Thus there will likely be obstacle in on the rise an effective solution Intrusion detection systems have been an area of active research for over fifteen years. Current commercial intrusion detection systems employ misuse detection. As such, they completely be short of the ability to detect new attacks. it is impossible to prevent security violation completely by using the exciting security technologies. Accordingly, Intrusion Detection is an significant component of network security.

VIII. REFERENCES

[1] Xu Wang, Beizhan Wang, Jing Huang," Cloud computing and its key techniques", ©2011 IEEE,pp 404-410
[2] Rajkumar Buyya1, Rajiv Ranjan2 and Rodrigo N. Calheiros," Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities", @csse.unimelb.edu.au, rajiv@unsw.edu.au, pp 1-11
[3] Sean Carlin, Kevin Curran," Cloud Computing Technologies",@ June 2012, International Journal of Cloud Computing and Services Science (IJ-CLOSER), Vol.1, No.2, ISSN: 2089-3337,http://iaesjournal.com/online/index.php/ IJ-CLOSER,pp 59-65
[4] Paul Martinaitis, Craig Patten and Andrew Wendelborn "Remote Interaction and Scheduling Aspects of Cloud Based Streams"2009 IEEE.

[5] Enda Barrett, Enda Howley, Jim Duggan" A Learning Architecture for Scheduling Workflow Applications in the Cloud" 2011 Ninth IEEE European Conference on Web Services978-0-7695-4536-3/11 $26.00 © 2011 IEEE.
[6] Boonyarith Saovapakhiran, George Michailidis†, Michael Devetsikiotis" Aggregated-DAG Scheduling for Job Flow Maximization in Heterogeneous Cloud Computing"2011 IEEE 978-1-4244-9268-8/11.
[7] Hsu Mon Kyi, Thinn Thu Naing" AN EFFICIENT APPROACH FOR VIRTUAL MACHINES SCHEDULING ON A PRIVATE CLOUD ENVIRONMENT"2011 IEEE 978-1-61284-159-5/11.
[8] V. Nelson, V. Uma" Semantic based Resource Provisioning and Scheduling in Inter-cloud Environment"2012 IEEE978-1-4673-1601-9/12 .
[9] Jinhua Hu, Jianhua Gu, Guofei Sun Tianhai Zhao "A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment"2010 IEEE 978-0-7695-4312-3/10.
[10] Shu-Ching, Wang, Kuo-Qin, Yan, Shun-Sheng, Wang, Ching-Wei, Chen" A Three-Phases Scheduling in a Hierarchical Cloud Computing Network" 2011 Third International Conference on Communications and Mobile Computing 978-0-7695-4357-4/11