ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

A STUDY ON IDS USING CLUSTERING TECHNIQUE IN DATA MINING

R.Bhavani [#1], K.Saranya[*2], Dr.G.Padmavathi[#3]

#R.Bhavani[1]
***K.Saranya[2]**
Research Scholar,
Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, India
bhavanisathes@gmail.com

#Dr.G.Padmavathi,[3]
Professor and Head,
Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, India

***Abstract*: Security and privacy of a system is most significant, to avoid intrusion in wireless network. Intrusion Detection System (IDS) plays significant role in network security as it perceives various attacks in network. Implementation of IDS is discerned between the traffic coming from network clients and the traffic begins from the attackers or intruders, in an endeavor to concurrently allay the problems of throughput, latency and security of the network. For this reason this study presents better approaches using data mining techniques. By pertaining Data Mining techniques on network traffic data is a hopeful result that facilitates to develop superior intrusion detection systems. Therefore this paper provides Study of various techniques of Intrusion detection system applied in the data mining with the help of clustering technique to show the effective detection of intrusion in network.**

***Keyword:* Data mining, IDS, wireless network, clustering and providing security.**

I INTRODUCTION

Data mining is the process of realize interesting knowledge from large amounts of data stored moreover in databases, data warehouses, or other information repositories. Data mining uses information from chronological data to scrutinize the product of a particular difficulty or circumstances that may arise. Data mining works to examine data accumulate in data warehouses that are used to accumulate that data that is being analyzed. That fastidious data may come from all parts of business, from the invention to the organization. Managers also use data mining to choose upon marketing advance for their product. They can use data to compare and contrast between contestants. Data mining interprets its data into real time examination that can be used to enlarge sales. Data mining has been implicated to scrutinize the useful information from huge level of data that are raucous, fuzzy and dynamic. While mining the data in network they were attackers shaped to lacerate the file from them without communicative the owner they hack the data because more and more computers receiving associated to public easily reached networks (e.g., the Internet), it is impractical for any computer system to be claimed confined to network intrusions. Since there is no ideal explanation to evade intrusions from happening, it is very important to be able to detect them at the first moment of happening and take actions to minimize the impending harm. For this prevention Intrusion detection is very imperative aspects of protecting the cyber connections from fanatic attack or from hackers.

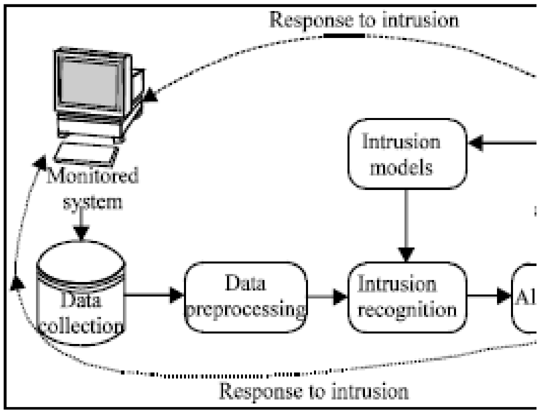

Figure 1: Architecture of IDS

IDS (Intrusion detection systems) system is the act of detecting and act in rejoinder to computer mishandling from the intruders. A system that executes mechanical intrusion detection is called an Intrusion Detection System (IDS). Intrusion hindrance technique such as firewall, riddling router policies fails to prevent much type of attacks. Therefore, no issue how secure we try to make our system, intrusion still happens and so they must be perceived. An IDS can be either host-based, if it monitors system calls or logs, or network-based if it monitors the stream of network packets. Modern IDSs are usually a combination of these three approaches. Another important variation is among systems that identify patterns of traffic or application data assumed to be malicious (misuse detection systems), systems that compare performance beside a ‘normal’ baseline (anomaly detection systems) and evaluate the behavior of objects with their associated object behavior (specification based techniques).

The Various Concepts of Intrusion Detection are

- **Prevention**: is to avoid intrusion by increasing the supposed risk of discovery and punishment
- **Discovery:** is to detect attacks and others security violation that are not prohibited by other security measures.
- **Anticipation:** is to detect and proactively deal with the activities that could designate that an attack is coming
- **False positive:** an event imperfectly identified by the IDS as being an intrusion when none has occurred.
- **False Negative:** actual intrusive action that IDS allows to pass as non-intrusive performance.
- **Subversion error:** when an intruder modifies the operation of the intrusion detector to force false negative to occur.

When a possible intrusion is exposed by an IDS, typical actions to execute would be logging relevant information to a file or database, generating an email alert, or produce a message to a pager or mobile phone. Formative what the probable intrusion actually is captivating some form of action to stop it or prevent it from occurrence again are frequently exterior the scope of intrusion detection. However, some forms of routine reaction have to represent the intruder behavior in wireless network, here with the help of Data Mining being studied and examine the clustering technique to perceive the intruder in the network and close by an Intrusion Detection System in data mining with different techniques. Misuse detection attempt to match known patterns of intrusion, while strangeness detection searches for leaving from normal behavior , this technique is used monitor the data from anomalous and normal one in network data and in measurement it detect during the behavior of the objects. A successful Intrusion detection system requires high detection rate, low false alarm rate as well as high correctness. This paper presents the evaluation on IDS and different Data mining techniques applied on IDS for the effective recognition of pattern for both malevolent and normal activities in network, which helps to develop secure information system.

II. LITERATURE REVIEW

From the author Mrutyunjaya Panda et al [5] study the Intrusion detection in the internet is an active area of research. A novel sequential information bottleneck (sIB) clustering algorithm has been proposed to build an efficient anomaly based network intrusion detection model. They have evaluate their planned method with other clustering algorithms like Farthest First, X-Means, DBSCAN, Filtered clusters, K-Means, and EM

(Expectation- Maximization) clustering in order to find the appropriateness of their proposed algorithm. A subset of KDDCup 1999 intrusion detection benchmark dataset has been used for the testing. Results show that the proposed method is resourceful in terms of detection exactness, low fake positive rate in association to the other existing methods. Their proposed approach supply better gratitude accuracy with rationally low false positive rate in decision to other accessible unsupervised clustering algorithms. This create the approach appropriate for building an efficient anomaly based network intrusion detection model. As apparent from the results none of the algorithms supply the best uncovering with zero false positive rates. Therefore, in the future research they shall investigate other data mining method with a view to enhance the detection accurateness as close as possible to 100% while safeguard a low false positive rate.

Manas ranjan patra and mrutyunjaya panda et al [6] analyze the current intrusion detection systems are signature based ones or machine learning based methods. Concerning the number of machine learning algorithms functional to KDD 99 cup, nothing have introduced a pre-model to make smaller the huge information quantity nearby in the different KDD 99 datasets. Clustering is an imperative task in mining evolving data streams. In addition the imperfect recollection and one-pass Constraints, the nature of embryonic data torrent entailes the following necessities for stream clustering: no announcement on the number of clusters, finding of clusters with arbitrary shape and ability to handle outliers. Conventional instance-based erudition methods can only be used to perceive known intrusions, since these procedure classify instances based on what they have erudite. In this paper, they suggest some clustering algorithms such as K-Means c-Means and Fuzzy for system intrusion detection. The experimental results attain by applying these algorithms to the KDD-99 data set express that they implement well in requisites of both accuracy and computation time.

Subaira.A.S and Anitha [7] study the information system, security has stay behind one solid line area for computers as well as networks. In information fortification, Intrusion Detection System (IDS) is used to preserve the data discretion, truthfulness and system accessibility from different types of attacks. Data mining is a inventive artifice applied to intrusion detection to establish a new delineate from the immense network data as well as it used to diminish the strain of the manual compilation of the normal and abnormal performance patterns. The DM clustering techniques, implement an intrusion detection system such as, disjointing methods, Hierarchical methods, Model based clustering technique and their an assortment of types.

From P. Uppuluri and R. Sekar [8] Specification-based intrusion detection, where actually specified program behavioral stipulation are used as a basis to notice attacks, have been predictable as a promising alternative that combine the potency of misuse gratitude (accurate detection of known attacks) and anomaly uncovering (ability to perceive novel attacks). However, the investigation of whether this promise can be realized in practice has stay following open. We answer this question in this paper, based on their accepting in building a specification-based intrusion detection system and conduct experiment with it. Their experiments incorporated the 1999 DARPA/AFRL online appraisal, as well as experimentation conduct using 1999 DARPA/ Lincoln Labs offline evaluation data. These experiments show that an successful specification-based ID can be developed with unpretentious efforts. They also show that the specification-based techniques live up to their guarantee of become aware of known as well as unknown attacks, while preserve a very low rate of false positives.

Nadya El Moussaid et al [9] The traditional Intrusion detection systems have been used elongated time ago, namely Anomaly-Based detection and Signature-based detection but have a lot of difficulty that limit their concert. Accordingly the main goal of this paper is to use data mining techniques including classification using clustering system to overpass these imperfection. This classification will be done by using k-means algorithm. Therefore they have improved k-means to triumph over its limits exclusively the cluster's number initialization. But they also present the development complete by initializing the cluster's numbers which is one of the limitations of this algorithm to make it more efficient in trounce some lacks of the established intrusion detection system, and also make it more intellectual and unsupervised.

III. CLUSTERING IN DATA MINING

Clustering analysis determine clusters of data objects that are similar in some understanding to one another. The members of a cluster are accompanying like each other than they are like members of other clusters. The goal of clustering examination is to find high-quality clusters such that the inter-cluster communication is low and the intra-cluster similarity is high.

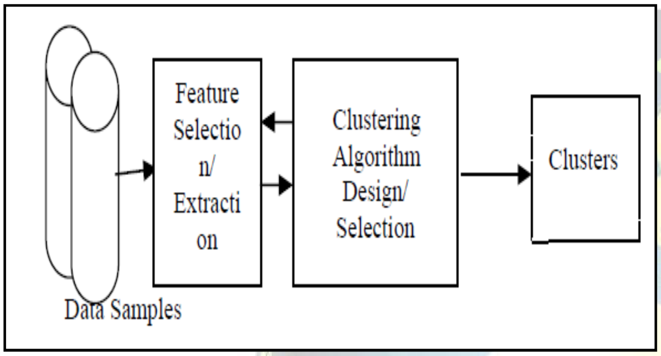
Feature Selectio n/ Extracti on
Clustering Algorithm Design/ Selection
Clusters
Data Samples

Figure 2: Clustering model

Clustering, like classification, is used to disjointing the data. Unlike classification, clustering models fragment data into groups that were not previously defined [10]. Clustering techniques in data mining are useful for a wide variety of real time submission production with large quantity of data. Clustering is an imperative task in mining developing data streams.

IV IDS USING CLUSTERING ANALYSIS:

Clustering is the classification of related objects into different groups, or additional precisely, the separation of a data into subsets (clusters), thus that the data in every subset (ideally) share some frequent trait-often proximity according to some defined distance determine. Clustering is a challenging field of explore as it can be used as a stand-alone tool to gain insight into the allotment of data, to observe the characteristics of each cluster, and to focus on a meticulous set of clusters for additional analysis. Clustering based intrusion detection systems categorized into three major groups:

- Signature detection systems
- Anomaly detection systems
- Specification-based Detection

a. Signature detection systems: In Signature detection, a model is trained with labeled data to distinguish the patterns of normal visits and an intrusion challenge is also called as misuse detection. Unlike the traditional knowledge base method, signatures of different types of intrusions are be trained mechanically, and they are much more prevailing than physically defined signatures in recording the delicate characteristics. Misuse detection has been shown to be very successful in detecting previously known attacks. However, since the misuse model is highly needy on the labeled data used in the training stage, its capabilities of detecting new intrusion types is limited. Dissimilar from misuse detection, anomaly detection first establishes a model of standard system behaviors, and anomaly measures are then eminent based on this cluster model. The implicit assumption is that any intrusive activity will be anomalous. In misuse detection advance, it defines abnormal system behavior at first, and then defines any other behavior, as normal behavior. It assumes that abnormal behavior and activity has a simple to define cluster model. It advances in the rapid of detection and low percentage of false alarm. However, it be unsuccessful in determine the non-pre-elected attacks in the feature library, so it cannot detect the abundant new attacks. The main advantage of misuse detection is that it can accurately detect known attacks from the cluster group, while its drawback is the inability to detect previously unseen attacks [11, 12]. With the help of the misuse analyses in clustering analyses it detect the intruder with the knowledge of the signature

b. Anomaly detection

Anomaly detection is able to detect recently emerging assault (if only the assumption still holds), but it also has some drawbacks. It may fail to detect some recognized attacks if the behaviors of them are not considerably different from what is considered to be normal. Anomaly detection notice anomalies in the data (i.e. data occurrence in the data that deviate from normal or regular ones) from the cluster object. It also allows us to detect new types of intrusions, because these new types will, by assumption, be deviations from the normal network usage. Moreover, the false alarm rate tends to be high when the data of some standard system behaviors are not occupied in the training phase. It refers to detect abnormal behavior of host or network from cluster group. It actually refers to storing features of user's usual behaviors hooked on database, then its compare user's present behavior with database. If there any deviation occurs, then it is said that the data tested is abnormal. Cluster Anomaly detection required the input to remain static for the network data. The patterns detected are called anomalies. Anomalies are also referred to as outliers. Anomaly detection, on the other hand, is capable of detecting novel attacks from the cluster group, but suffers from a high rate of false alarms. This occurs primarily because previously unseen (yet legitimate) system behaviors are also recognized as anomalies, and hence flagged as potential intrusions [5, 11, 13].

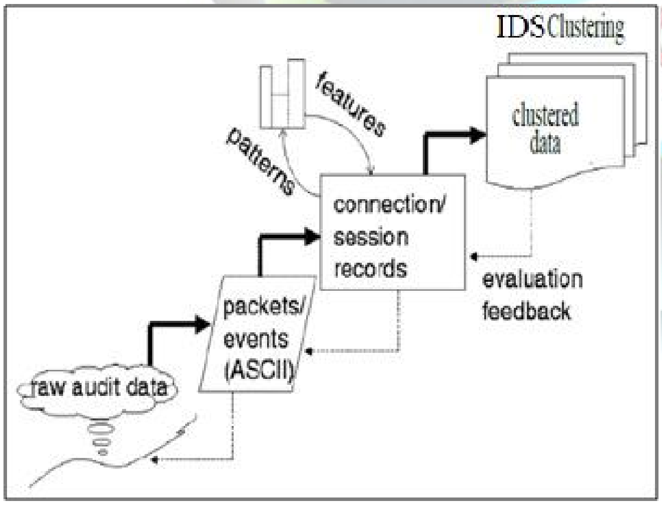

Figure 3: Clustering analyses using IDS

c. Specification Detection

Specification-based intrusion detection, where manually specified program behavioral specifications and used as a basis to detect attacks. Clustering Specification-based techniques have been used to improve more security in network from the intruders. The idea is to use traces, ordered sequences of implementation events, to identify the intended behaviors of simultaneous programs in distributed system. A requirement explain valid process sequences of the implementation of one or more programs, collectively called a (monitored) subject. A succession of operations performed by the subject that does not conform to the specification is considered a security violation. Specification based detection it is worked with the help of parameterized of cluster group with respect to system calls as well as their arguments.. It automatically generates the specification of a program by deriving an abstract cluster model. This model adopts different ways to represent the possible system call traces according to the control flow graph of clustering analysis. Attractive features of this approach are that it has the potential to detect unknown patterns of attacks from the clustering object and it has no false alerts, although it may miss some attacks [14, 15]. Specifications obtained from the previous steps are customized to accommodate difference in operating systems, and as well any site specific security policies. Specifications increase the effectiveness of the system at the cost of increased specification development effort. The advantage of this approach is, it should be able to detect some new types of attacks with the knowledge of cluster that intruders will invent in the future.

Technique	**Advantage**	**Disadvantage**
Misuse detection	Accurately detect known attacks & low percentage of false alarm	Inability to detect previously unseen attacks.
Anomaly detection	Detect new types of intrusions	False alarm result

Specification based Detection	Generates Automatically	Increasing in cost

Table 1: Comparison analyses

This table shows the analyses and comparison of the IDS techniques with the help of Clustering in Data mining.

V IDS USING CLUSTERING ALGORITHM

a.K-Means Clustering algorithm

K-Means clustering algorithm is simplest and widely used clustering technique. In this algorithm, number of clusters K is specified by user means classifies instances into predefined number of cluster. The first step of K-Means clustering is to choose k instances as a center of clusters. Next assign each instances of dataset to nearest cluster. For instance assignment, measure the distance between centroid and each instances using Euclidean distance and according to minimum distance assign each and every data points into cluster. K – Means algorithm takes less execution time, when it applied on small dataset. When the data point increases to maximum then it takes maximum execution time. It is fast iterative algorithm but it is sensitive to outlier and noise. This is used to find the attackers in wireless network by the instance of the center point. Then it assigns each attacker to the nearest point or node then it measure the attackers according to the minimum distance and it report the intruder to the server.

b. K-Medoids clustering algorithm

K-Medoids is clustering by partitioning algorithm as like as K-means algorithm. The most centrally situated instance in a cluster is considered as centroid in place of taking mean value of the objects in K-Means clustering. This centrally located object is called reference point and medoid. It minimizes the distance between centroid and data spoints means minimize the squared error. KMedoids algorithm performs better than K-Means algorithm when the number of data points increases to maximum. It is robust in presence of noise and outlier because medoid is less influenced by outliers, but processing is more expensive. This is used to find the attackers in wireless network by the instance of the center point. It act like the K-Means cluster algorithm and then it assigns each attacker to the nearest point or node with the help of IDS and also it choose the references object also to find the intruder.

c. Expectation Maximization (EM)

The EM algorithm is used to searches the result for a maximum likelihood. It estimates the expected values of the hidden variables. The hidden variables are the parameters of the model. In this case, we use mixture of Gaussians; hence the hidden variables are the mean and standard deviation for each Gaussian distribution. We start with an initial estimate of those parameters and iteratively run the algorithm to find the maximum likelihood (ML) for our estimates. The reason we are using EM is to fit the data better, so that clusters are compact and far from other clusters, since we initially estimate the parameters and iterate to find the ML for those parameters. With the help of the EM algorithm it finds out the hidden attackers in the network. It iteratively process of IDS the activity and analyses the attackers in the network, it provide a results in maximum expected attackers in the network. From the above three algorithm EM provide more accuracy rate of intruders.

Algorithm	**Advantage**	**Disadvantage**
K-Means algorithm	Fast iterative to find attackers	Sensitive to outlier and noise
K-Medoids algorithm	It is robust in presence of noise and outlier, Shows maximum attacks, less false alarm rate	Processing provide more expensive

Expectation Maximization (EM)	Find hidden attacks, more accuracy, less false alarm rate, higher detection rate	Expensive

Table 2: Compare analyses of Clustering algorithm in IDS.

VI CONCLUSION

In this paper we study and analyze the various concept of the intrusion detection which falls n the clustering data mining technique. Mostly in this current world they where many hackers are raised to scrutinize the authorized data from the authorized user, with the increasing growth in network attackers also increased. To prevent the data from those attackers is the most vital one, here we present the survey of different analyses of various technique to avoid and prevent from the intruder, clustering analyses of data mining helps to solve these problem, we present the three different approach on clustering technique; misuse, anomaly and specification detection, specification prove the best result when compared to the other technique, specification of clustering act important role to find the intruders.

VII REFERENCES

[1]. Peng Ning, "Intrusion Detection Techniques"

[2]. http://www.slideshare.net/Tommy96/data-mining-techniques-for-network-intrusion-detection-systems.

[3]. Abhaya, Kaushal Kumar, Ranjeeta Jha, Sumaiya Afroz, "Data Mining Techniques for Intrusion Detection: A Review".

[4]. Subaira.A.S1, Anitha, "A Study of Network Intrusion Detection by Applying Clustering Techniques"

[5]. Mrutyunjaya Panda et al, "A Novel Classification via Clustering Method for Anomaly Based Network Intrusion Detection System"

[6]. Manas ranjan patra and mrutyunjaya panda, "Some clustering algorithms to enhance the performance of the network intrusion detection system"

[7]. Subaira.A.S1, Anitha, "A Study of Network Intrusion Detection by Applying Clustering Techniques"

[8]. P. Uppuluri and R. Sekar, "Experiences with Specification-based Intrusion Detection?"

[9]. Nadya El Moussaid, Ahmed Toumanari, Maryam Elazhari, "Intrusion Detection Based On Clustering Algorithm".

[10]. Qiang Wang, "A Clustering Algorithm for Intrusion Detection"

[11]. G.Keerthana and Dr.G.Padmavathi, "A Study on Sinkhole Attack Detection using Swarm Intelligence Techniques for Wireless Sensor Networks"

[12]. Nadya El MOUSSAID et al, "Intrusion Detection Based On Clustering Algorithm"

[13]. Wenke Lee, Salvatore J. Stolfo , Philip K. Chan, "Real Time Data Mining-based Intrusion Detection".

[14]. Jian Pei, Shambhu J. Upadhyaya, "Data Mining for Intrusion Detection – Techniques, Applications and Systems".

[15]. R. Sekar, et al, "Specificationbased Anomaly Detection: A New Approach for Detecting Network Intrusions".

AUTHOR'S BIOGRAPHY

R.Bhavani received her M.Sc Computer Science degree in 2015 from Kongunadu college of Arts & Science, Coimbatore. She is pursuing her M.Phil at Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore. Her areas of interest are Data Mining, Security.

Dr.G.Padmavathi is the Professor and Head of computer science Department of Avinashilingam Institute for Home Science and Higher Education for Women University, Coimbatore. She has 27 years of teaching experience and one year of industrial experience. Her areas of interest include Real Time Communication, Wireless Communication, Network Security and Cryptography. She has significant number of publications in peer reviewed International and National Journals. Life member of CSI, ISTE, WSEAS, AACE and ACRS.