ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 1, Issue 8, October 2013**

# A Study of Network Intrusion Detection by Applying Clustering Techniques

Subaira.A.S[1], Anitha.P[2]

PG Scholar, Department of CSE, Dr.N.G.P.Institute of Technology, Coimbatore, India[1]

Assistant Professor, Department of CSE, Dr.N.G.P.Institute of Technology, Coimbatore, India[2]

**ABSTRACT***:* In information system, security has remained one hard line area for computers as well as networks. In information protection, Intrusion Detection System (IDS) is used to safeguard the data confidentiality, integrity and system availability from various types of attacks. Data mining is an efficient artifice applied to intrusion detection to ascertain a new outline from the massive network data as well as it used to reduce the strain of the manual compilations of the normal and abnormal behavior patterns. This piece of writing reviews the present state of data mining clustering techniques to implement an intrusion detection system such as, Partitioning methods, Hierarchical methods, Model based clustering methods and their various types.

***Keywords***: Clustering, Intrusion Detection System, Data Mining, Anomaly Detection, Misuse Detection

## I. INTRODUCTION

In the era of information society, as network-based computer systems play fundamental roles, they have become the target for intrusions by the attackers and criminals. Intrusion prevention technique such as firewalls, user authentication, information protection and data encryption have failed to completely shield networks and systems behavior from the growing and sophisticated attacks and malwares. To defend the computers and networks from various cyber attacks and viruses the Intrusion Detection Systems (IDS) are designed. An IDS is a mechanism that monitors network or system actions for malicious activities and produces reports to a management station [1].

As a significant application area of data mining is intrusion detection based on data mining algorithms, aims to solve the troubles of analyzing enormous volumes of data [9]. IDSs build efficient clustering and classification models to distinguish normal behaviors from abnormal behaviors using data mining techniques. This study makes foundation in this field of research and exploration and implements intrusion detection model system based on data mining technology.

## II. TRADITIONAL INTRUSION DETECTION

There are two types of traditional intrusion detection system

### A. ANOMALY DETECTION

It refers to detect abnormal behaviour of host or network. It actually refers to storing features of user's usual behaviours hooked on database, then its compare user's present behaviour with database. If there any deviation occurs, then it is said that the data tested is abnormal [2]. The patterns detected are called anomalies. Anomalies are also referred to as outliers.

### B. MISUSE DETECTION

In misuse detection approach, it defines abnormal system behaviour at first, and then defines any other behaviour, as normal behaviour. It assumes that abnormal behaviour and activity has a simple to define model. It advances in the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 1, Issue 8, October 2013**

rapid of detection and low percentage of false alarm. However, it fails in discovering the non-pre-elected attacks in the feature library, so it cannot detect the abundant new attacks [3].

IDS provide the following security functions

A. DATA CONFIDENTIALITY

It checks whether the information stored on a system is protected against unconstitutional access. Since systems are sometimes used to manage sensitive information, data confidentiality is often a gauge of the ability of the system to protect its data [4].

B. DATA INTEGRITY

It refers to maintaining and assuring the correctness and consistency of data over its entire life-cycle. No corruption or data loss is acknowledged either from random events or malicious activity.

C. DATA AVAILABILITY

The network should be tough to Denial of Service attacks.

Intrusion detection system based on sources of audit information it can be divided into 3 subcategories

A. HOST BASED IDS

It refers to intrusion detection that takes place on a single host system. It gets audit data from host audit trails and monitors activities such as integrity of system, file changes, host based network traffics, and system logs. If there is any unlawful change or movement is detected, it alerts the user by a pop-up menu and informs to the central management server. Central management server blocks the movement or a combination of the above three [5]. The judgment should be based on the strategy that is installed on the local system.

B. NETWORK BASED IDS

It is used to supervise and investigate network transfer to protect a system from network-based threats. It tries to detect malicious activities such as denial-of-service (Dos) attacks and network traffic attacks. Network based IDS includes a number of sensors to monitors packet traffic, one or more servers for network management functions, and one or more management relieves for the human interface [6].

C. HYBRID INTRUSION DETECTION

The recent development in intrusion detection is to combine both types host-based and network-based IDS to design hybrid systems. Hybrid intrusion detection system has flexibility and it increases the security level. It combines IDS sensor locations and reports attacks are aimed at particular segments or entire network [7].

## III. TYPES OF ATTACKS

A. DOS ATTACK

A denial-of-service attack or distributed denial-of-service attack is an effort to make a computer resource out of stock to its Intended users [32]. In this type of attack it slow down the system or shut down the system so it disrupt the service and deny the legitimate authorized user. Due to this attack high network traffic occurs [10].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 1, Issue 8, October 2013**

B. USER TO ROOT ATTACK (U2R)

In this type of attack the attacker starts with user level like taking down the password, dictionary attack and finally attacker achieves root to access the system.

C. PROBING

In this type of attack an attacker examines a network to gather information or discover well-known vulnerabilities. An attacker who has a record, of which machines and services are accessible on a known network, can make use of this information to look for delicate points.

D. REMOTE TO USER ATTACK (R2U)

In this type of attack an attacker have the capability to send packet to a machine over a network but does not have an account on that machine, make use of some vulnerability to achieve local access as a user of that machine.

E. EAVESDROPPING ATTACK

Eavesdropping is a network layer attack consisting of capturing packets from the network transmitted by others' computers and reading the sensitive information like passwords, session tokens, or any kind of confidential information.

F. MAN-IN-THE-MIDDLE ATTACK

In this the attacker makes independent connections with the victims and relays messages between them and making them believe that they are talking directly to each other over a private connection, but the fact is entire conversation is controlled by the attacker.

## IV. DRAWBACKS OF IDS

Intrusion Detection Systems (IDS) have become an important component in security infrastructures as they permit networks administrators to identify policy variations. These policy violations range from outside attackers trying to gain unconstitutional access to intruders abusing their access. Current IDS have a number of considerable drawbacks

FALSE POSITIVES**:** A major problem is the amount of false positives IDS will produce. Developing distinctive signatures is a complicated task. It is much trickier to pick out a legitimate intrusion attempt if a signature also alerts regularly on valid network activity.

FALSE NEGATIVES**:** In these IDS does not generate an alert when an intrusion is actually taking place. It simply put if a signature has not been written for a particular exploit there is a tremendously good chance that the IDS will not detect it**.**

| | **Intrusion** | **Normal** |
|---|---|---|
| **Intrusion** | True Positives(TP) | False Negatives(FN) |
| **Normal** | False Positives(FP) | True Negatives(TN) |

**Table 1: Performance Measure**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 1, Issue 8, October 2013**

To identify the Accuracy Rate and False Positive rate the following two formulas can be used

$$\text{Accuracy Rate} = \frac{FP+FN}{TP+TN+FP+FN}$$

$$\text{False Positive} = \frac{FP}{TP+FP}$$

## V.DATA MINING ASSISTS FOR INTRUSION DETECTION

The central theme of intrusion detection using data mining approach is to detect the security violations in information system. Data mining can process large amount of data and it discovers hidden and ignored information. To detect the intrusion, data mining consists of process like, classification, clustering, and regression [8]. In this work we have focused on Clustering techniques. It monitors the information system and raises alarms when security violations are founded.

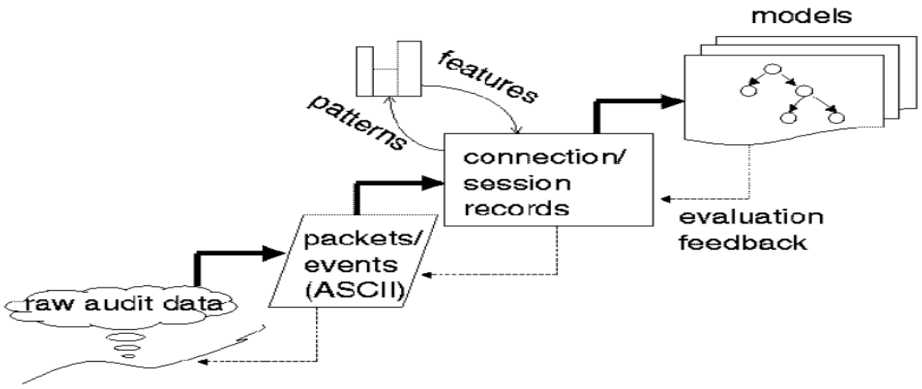


**Figure 1[19]: The Data mining method of Building ID Models**

## VI. CLUSTERING TECHNIQUES

### A. PARTITIONING METHODS

1. K-MEANS CLUSTERING

The widely used clustering algorithm is K-Means which was proposed by James MacQueen.K-Means is most powerful clustering algorithm that is used in machine learning which can be used to recognize groups of similar instances, substance, objects, and points automatically in data training. The algorithm categorizes instances to a pre-defined number of clusters which is specified by the user. The first important step is to choose an input parameter, k, and partition the n objects into k clusters and it randomly selects k objects. Second step mainly used to read each instance from the data set and it assigns it to the nearest cluster. In IDS this algorithm initially chooses a cluster head, cluster header is chosen randomly. After choosing clustering head it start clustering process and similar data is assigned to its neighbors. So it accurately classifies the normal and abnormal data. The most commonly used method to measure the distance between instance and the centroid is Euclidian distance. After every insertion of instance the cluster centroids are recalculated. This process is recapitulated until no more changes are made. That is, square-error criterion is used, defined as

$$E=\sum_{i=1}^{k}\sum_{p\in Ci}|p-mi|^2,$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 1, Issue 8, October 2013**

Where E is the sum of the square error for all objects in the data set; p is the point in space representing given object; and $m_i$ is the mean of cluster $C_i$(p and $m_i$ are multidimensional)[21]

2. K-MEDOIDS

K-Medoids attempts to minimize the distance between points and its centroid. This clustering algorithm is similar to k-Means. A medoid is a data point which acts as an exemplar for all other data points in the cluster. The k-Means algorithm is very sensitive to outliers because if there is an object with a very large value, the data distribution may be biased or distorted [11]. In this case, k-Medoids is more robust to noise and outliers because in this algorithm the partitioning method is performed based on the principle of minimizing the sum of dissimilarities between each object in a cluster [11]. Typically, the absolute-error is used, defined as

$$E=\sum_{j=1}^{k}\sum_{p\in Cj}|p-oj|,$$

Where E is the sum of the absolute error for all objects in the data set; p is the point in space representing a given object in cluster Cj; and $o_{j\ is}$ the representative object of $C_j$ [21]

3. OUTLIER DETECTION ALGORITHMS

Outlier detection is a technique to find patterns in data that do not conform to expected behavior [12]. Most of the clustering algorithms do not assign all points to clusters but account for noise objects, in other words clustering algorithms are optimized to find clusters rather than outliers. Outlier detection algorithms look for outliers by applying one of the clustering algorithms and retrieve the noise set; therefore the performance of outlier detection algorithms depends on how good the clustering algorithm captures the structure of clusters. The distance-based outlier detection approach, which is based on the Nearest Neighbor algorithm was first introduced by Ng et al [13] and implements a well-defined distance metric to detect outliers, the greater the distance of the object to its neighbour,the more likely it is to be an outlier[14]. This method calculates the distance between each pair of objects using a nested loop (NL) algorithm and then the objects which are far away from the majority are signed as outliers [13].

**B.HIERARCHICAL METHODS**

1. HIERARCHICAL CLUSTERING ALGORITHMS

These algorithm works by grouping of data items into a tree of clusters. It can be classified as either agglomerative or divisive; depending upon the hierarchical decomposition is formed in a bottom up (merging) or top down (splitting) fashion.

2. AGGLOMERATIVE HIERARCHICAL CLUSTERING

It is based on bottom up strategy. This algorithm mainly used to make each sample as a separate cluster and these clusters merged with a least distance to get larger until the termination condition is satisfied or a single cluster is left.

3. DIVISIVE HIERARCHICAL CLUSTERING

It is based top down strategy. It mainly performs the operation opposite to that of agglomerative hierarchical clustering by considering all objects in one cluster. It separates the clusters into smaller pieces, until each object has its cluster of its own or it satisfies the termination condition.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

4. BIRCH: BALANCED ITERATIVE REDUCING AND CLUSTERING USING HIERARCHIES

BIRCH is a clustering techniques designed for handling large amount of numerical data by integrating hierarchical clustering and other clustering methods such as iterative partitioning. BIRCH is also the first clustering algorithm proposed in the database area to handle noise effectively. It includes two concepts that is clustering feature (CF) and clustering feature tree (CF tree), which produces the cluster representation. While handling the large amount of data it produces the good speed and scalability.

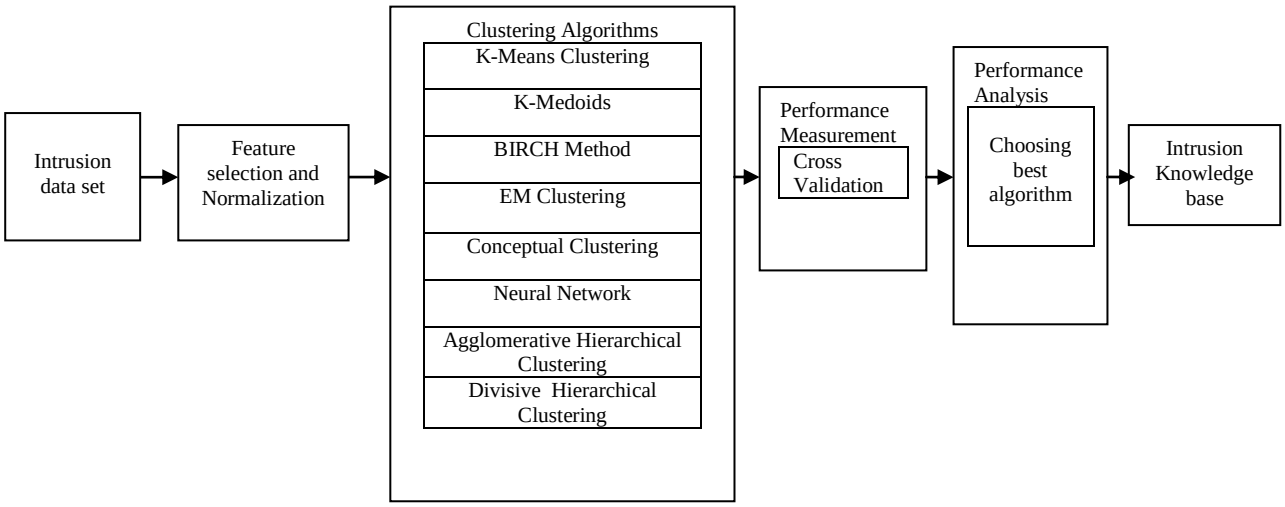


**Figure 2: Intrusion Detection System Using Clustering Algorithms**

**C.MODEL BASED CLUSTERING METHODS**

1. EM CLUSTERING

Expectation Maximization (EM) clustering is a variant of k-Means clustering and is widely used for density estimation of data points in an unsupervised clustering [15]. EM algorithm is used to discover the framework which maximizes the possibility of the data, assume that the data is generated from k normal distributions. In IDS initially this algorithm discovers the framework and identifies the intruded data. The algorithm learns both the means and the covariance of the normal distributions. This method requires several inputs which are the data set, the total number of clusters, the maximum error tolerance and the maximum number of iteration. The EM can be divided into two important steps which are Expectation (E-step) and Maximization (M-step). The goal of E-steps is to calculate the expectation of the likelihood (the cluster probabilities) for each instance in the dataset and then re-label the instances based on their probability estimations.

$$P(x_i \in C_k) = p(C_k|x_i) = \frac{p(Ck)\, p(xi|Ck)}{p(xi)}$$

Where $p(x_i|C_k) = N(m_k, E_k(x_i))$ follows normal distribution around mean,$m_k$ with expectation Ek.

The M-step is used to re-estimate the parameters values from the E-step results. The outputs of M-step (the parameters values) are then used as inputs for the following E-step.

$$m_k = \frac{1}{n}\sum_{i=1}^{n} \frac{xiP(xi \in Ck)}{\sum j\, P(xi \in Cj)}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 1, Issue 8, October 2013**

These two processes are performed iteratively until the results convergence. The mathematical formulas of EM clustering are described in [15] [16] and the pseudo codes can be found in [16].

2. CONCEPTUAL CLUSTERING

It is a machine learning clustering techniques, for a given set of unlabeled objects it produces a classification scheme over the objects. In IDS it classifies the unlabeled object and produces a good classification scheme. Conceptual clustering also finds the characteristic descriptions for each group. Each group represents a concept or class. It includes two step processes: First it performs clustering and second one is characterization. So in IDS after performing clustering it's characterized each data and produces the report to management station. For an incremental conceptual clustering COBWEB is a well-liked method. COBWEB creates hierarchical clustering in the form of classification tree.

3. NEURAL NETWORK APPROACH

Neural Network was traditionally used to refer a network or biological neurons. [17]In IDSs neural network has been used for both anomaly and misuse intrusion detection. In anomaly intrusion detection the neural networks were modeled to recognize statistically significant variations from the user's recognized behavior also identify the typical characteristics of system users. In misuse intrusion detection the neural network would collect data from the network stream and analyze the data for instances of misuse [18].In neural network the misuse intrusion detection can be implemented in two ways. The first approach incorporates the neural network component into an existing system or customized expert system. This method uses the neural network to sort the incoming data for suspicious events and forward them to the existing and expert system. This improves the efficiency of the detection system. The second method uses the standalone misuse detection system. This system receives data from the network stream and analyzes it for misuse intrusion. It has the ability to learn the characteristics of misuse attacks and identify instances that are unlike any which have been observed before by the network. It has high degree of accuracy to recognize known suspicious events. Generally, it is used to learn complex non linear input-output relationships [20].

## IV.PERFORMANCE ANALYSIS

| Algorithm | Accuracy | False Positive |
|---|---|---|
| K-Means Clustering | 75.41% | 22.95% |
| K-Medoids | 76.72% | 21.83% |
| Outlier Detection Algorithm | 80.14% | 21.83% |
| BIRCH Method | 76.82% | 18.16% |
| EM Clustering | 78.06% | 20.74% |
| Conceptual Clustering | 76.41% | 23.16% |
| Neural Network | 79.17% | 22.13% |
| Agglomerative Hierarchical Clustering | 74.20% | 21.46% |
| Divisive Hierarchical Clustering | 73.24% | 22.37% |

**Table 2: The Algorithm result using Clustering techniques**

## V. CONCLUSION

In this paper, many data mining techniques have been proposed to improve the detection rate of Intrusion Detection System. In future, we planned to combine more than one clustering technique because different clustering algorithm have different knowledge to solve the problem so combining more than one data clustering algorithm is used to remove

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 1, Issue 8, October 2013**

the demerits of one another and a number of trained classifier lead to a superior performance than any single classifier. These techniques provide better performance in Intrusion Detection accuracy rate, faster running time and detecting the false positive rate. To fragment a complex problem into sub problems for which the solutions obtained are simpler to realize, execute, supervise and update.

## REFERENCES

[1] W. Lee, S.J. Stolfo, K.W. Mok, "A data mining framework for building intrusion detection models", in: Proceedings of IEEE Symposium on Security and Privacy, pp. 120–132, 1999.
[2] J .X. Huang, J.Miao, Ben He, "High performance query expansion using adaptive co –training", Information Processing & Management, pp. 441–453, 2012.
[3] S. Axelsson," Research in intrusion detection systems a survey", Chalmers University of Technology, Goteborg, Sweden, in: Tech. Rep.TR98-17, 2000.
[4] T.F. Lunt, "A survey of intrusion detection techniques", Computers and Security, pp. 405–418.1993.
[5] S. Freeman, J. Branch, "Host-based intrusion detection using user signatures", in: Proceedings of the Research Conference RPI., 2002.
[6] D. Marchette, "A statistical method for profiling network traffic", in: Proceedings of Workshop on Intrusion Detection and Network Monitoring, pp. 119–128, 1999.
[7] Crosbie. M, E. H. Spafford,"Active defense of computer system agents", Technical Report CSD-TR- 95-008, Purdue Univ.West Lafayette, IN, 1995.
[8] T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: an efficient data clustering method for very large databases", in: Proceedings of SIGMOD, ACM, pp. 103–114, 1996.
[9] V.Vapnik, "The Nature of Statistical Learning Theory", Springer, 1999.
[10] T.A. Longstaff, J.T. Ellis, S.V. Hernan, "Security of the Internet", in: F. Froehlich, A. Kent (Eds.), the Froehlich/Kent Encyclopedia of Telecommunications.Vol.15, pp. 231–254, 1998.
[11] Velmurugan, T., Santhanam. T ," Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points". Journal of Computer Science , pp 363–368, 2003.
[12] Chandola, V," Anomaly detection: A survey".ACM Computer society 41 (3) 1–58, 1998.
[13] Knorr,"Finding Intentional Knowledge of Distance-Based Outliers".In Proceedings of the 25thInternational Conference on Very Large Data Bases, 211–222, 2009.
[14] Orair," Distance-based outlier detection: consolidation and renewed bearing". In: Proceedings VLDBEndow 3, 1-2, 1469–1480, 2007.
[15] Seetha,"Unsupervised Learning Algorithm for Color Texture Segmentation Based Multiscale Image Fusion", European Journal of Scientific Research 67 (4) 506–511, 2006.
[16] Lu, W., Tong, H... "Detecting Network Anomalies Using CUSUM and EM Clustering". In: Proceedings of the 4thInternational Symposium on Advances in Computation and Intelligence, p. 297–308, 2009.
[17] J. Ryan, M.-J. Lin, R. Miikkulainen,"Intrusion detection with neural networks", in: Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection and Task Management, pp. 92–97,1997.
[18] D.E. Denning,"An intrusion-detection model", IEEE Transactions on Software Engineering,pp. 222–232, 1987.
[19]http://wenke.gtisc.gatech.edu/project/image004.gif
[20].H. Brahmi, I. Brahmi, S.B. Yahia, "OMC-IDS: at the cross-roads of OLAP mining and intrusion detection", in: Advances in Knowledge Discovery and Data Mining, in: LNCS, vol. 7302, pp. 13–24, 2012.
[21]J.Han, M.Kamber"Data Mining Concepts and Techniques", in: ELSEVIER, Second Edition, 2006.