



# Anomaly Detection using Clusters and Proximities Measures

**Sumathy Murugan**

*Research and Development Centre  
Bharathiyar University  
Coimbatore  
sumathymurugan@gmail.com*

**M Sundara Rajan**

*Department of Computer Science  
Government Arts College, Nandanam  
Chennai  
Drmsrajan23@yahoo.com*

**Abstract-** In an increasing number of security issues, intruder detection system are used to detect an insecure network attacks. There are so many attacks, in real time process; to detect it some of IDS system is used for filtering such data packets. This paper analysis the anomaly based intrusion detection techniques. AIDS is a system for detecting intrusions, type of attacks that falls out of normal process system activity and classifying it as either normal or anomalous. Anomaly detection searches for an unusual cases based on behavior analysis deviations. It quickly detects the attack in data analysis process by clustering. A StepWiseClustering (SWC) algorithm is used to detect the attack in unusual cases.

Keywords- IDS, HA Clustering Algorithm, Proximities Measure, Anomaly Detection

## I. INTRODUCTION

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices [1, 3]. Intrusion Detection Systems (IDS) are security tools that behave like the other measures such as antivirus software, internet security, and firewalls, are intended to strengthen the security of information and communication systems [5].

CIDF Common Intrusion Detection Framework, a working group created by DARPA in 1998 and integrated within IETF in 2000 and adopted the new acronym IDWG Intrusion Detection Working Group, defined a general IDS architecture based on the consideration of four types of functional modules [6].

Encryption mechanisms are designed to protect data against passive attacks. Hence, one can say that there is a need to design mechanisms that are capable enough of detecting and preventing multiple security attacks in network [2]. An Intrusion Detection System (IDS) is one possible solution to it. An intrusion is basically any sort of unlawful activity which is carried out by attackers to harm network resources or sensor nodes. An IDS is a mechanism to detect such unlawful or malicious activities [7]. The primary functions of IDS are to monitor users' activities and network behavior at different layers. IDS can operate in many modes, for example, stand-alone operation and cooperative cluster based operation [8]. A standalone IDS operates on every node to detect unwanted activities. Cooperative cluster based IDS are mostly distributed in nature in which every node monitors its neighbors and surrounding nodes activities and operation; in case of any malicious activity detection, the cluster head is informed.

## II. INTRUDERS DETECTION TYPES

Signature based detection is a pattern that corresponds to a known threat. Signature-based detection is the process of comparing signatures against observed events to identify possible incidents. Anomaly based detection IDS that looks at network traffic and detects data that is incorrect, not valid, or generally abnormal is called anomaly-based detection [1]. Hybrid IDS are a combination of both anomaly-based and signature-based approaches. Hybrid mechanisms usually contain two detection modules; that is, one module is responsible of detecting well-known attacks using signatures, while the other is responsible for detecting and learning normal and malicious patterns or monitor network behavior deviation from normal profile. Hybrid IDSs are more accurate in terms of attack detection with less number of false positives [5, 4].

### A. Signature-Based Detection Vs Anomaly-Based Detection

Intrusion detection systems are classified as either signature-based or anomaly-based. Signature based detectors or misuse-based find known patterns, or signatures, within the analyzed data. For this purpose, a signature database corresponding to known attacks is specified a priori. On the other hand, anomaly-based detectors attempt to estimate the "normal" behavior and generate an anomaly alarm whenever the deviation

between a given observation at that time and the normal behavior exceeds a predefined threshold. Another possibility is to model the “abnormal” behavior of the system and to raise an alarm when the difference between the observed behavior and the expected one falls below a given threshold. The main differences between these methodologies are inherent in the concepts of “attack” and “anomaly” [4].

- An attack can be defined as “a sequence of operations that puts the security of a system at risk”.
- An anomaly is just “an event that is suspicious from the perspective of security”.

### III. ANOMALY - BASED INTRUSION DETECTION SYSTEM (AIDS)

Anomaly-based IDS monitors network activities and classifies them as either normal or malicious using heuristic approach. Most of anomaly-based IDSs identify intrusions using threshold values; that is, any activity below a threshold is normal, while any condition above a threshold is classified as an intrusion. The main advantage of anomaly-based IDS is its capability to detect new and unknown attacks; however sometimes it fails to detect even well-known security attacks. Many anomaly-based IDSs have been proposed so far [10]. This is especially true for larger networks and, with high bandwidth connections, it is therefore necessary to install the anomaly sensors closer to the servers and network that are being monitored. Anomaly-based intrusion detection triggers an alarm on the IDS when some type of unusual behavior occurs on your network that can be any event, state, content, or behavior that is considered to be abnormal by a pre-defined standard.

Anything that deviates from this baseline of “normal” behavior will be flagged and logged as anomalous. “Normal” behavior can be programmed into the system based on offline learning and research or the system can learn the “normal” behavior online while processing the network traffic. Anomaly based intrusion detection, on the other hand, takes a more generalized approach when looking for and detecting threats to your network [5, 9]. A baseline of “normal” behavior is developed, and when an event falls outside that norm, it is flagged and logged. The behavior is a characterization of the state of the protected system, which is both reflective of the system health and sensitive to attacks. In this context, an anomaly-based method of intrusion detection has the potential to detect new or unknown attacks. Like the signature-based method, however, anomaly-based intrusion detection also relies on information that tells it what is normal and what isn't. This is called a profile, and it is key to an effective anomaly-based intrusion detection system.

TABLE 1. COMPARISON OF DIFFERENT IDS

Characteristics	Signature based IDS	Anomaly based IDS	Hybrid IDS
Detection rate	Medium	Medium	High
False alarm	Medium	Medium	Low
Computation	Low	Low	Medium
Energy consumption	Low	Low	Medium
Attack detection	Few	Few	More
Multilayer attack detections	No	No	No
Strength	Detects all those attacks having signatures	Capable of detecting new attacks	Can detect both existing and new attacks
Weakness	Cannot detect new attacks	Misses well known attack	Requires more computation and resources

### IV. ASSESSMENT OF ANOMALY IDS

Two key aspects concern the evaluation, and comparison, of the performance of alternative intrusion detection approaches: these are the efficiency of the detection process, and the cost involved in the operation. Without the importance of the cost, at this point the efficiency aspect must be emphasized.

The main benefit of anomaly-based detection techniques is their potential to detect previously unseen intrusion events. However, and despite the likely inaccuracy in formal signature specifications, the rate of false positives (or FP), events erroneously classified as attacks in anomaly-based systems is usually higher due to the ever changing nature of networks, applications and exploits.

#### A. False Positives and Negatives

Four situations exist in this context, corresponding to the relation between the result of the detection for an analyzed event either “normal” or “intrusion” and its actual nature either “innocuous” or “malicious” [1]. These situations are:

- False Positive (FP), if the analyzed event is innocuous (or “clean”) from the perspective of security, but it is classified as malicious;
- True Positive (TP), If the analyzed event is correctly classified as intrusion/malicious;
- False Negative (FN), if the analyzed event is malicious but it is classified as normal/innocuous; and
- True Negative (TN), if the analyzed event is correctly classified as normal/innocuous.

False positive rate is measured over normal data items. Suppose that  $m$  normal data items are measured and  $n$  of them are identified as abnormal.

*False positive rate is defined as  $n/m$ .*

Detection rate is measured over abnormal data items. Suppose that  $m$  abnormal data items are measured, and  $n$  of them are detected.

*Detection rate is defined as  $n/m$ .*

## V. RELATED WORKS

Anomaly-based network intrusion detection techniques are a valuable technology to protect target systems and networks against malicious activities, classification of the anomaly detection techniques according to the nature of the processing involved in the “behavioral” model considered. Anomaly detection systems, a subset of intrusion detection systems, model the normal system/network behavior which enables them to be extremely effective in finding and foiling both known as well as unknown or “zero day” attacks. While anomaly detection systems are attractive conceptually, a host of technological problems need to be overcome before they can be widely adopted.

[A.Patcha JM park 2007] We present and demonstrate the use of a general-purpose hierarchical multitier multi window statistical anomaly detection technology and system that operates automatically, adaptively, and proactively, and can be applied to various networking technologies, including both wired and wireless ad hoc networks. This method uses statistical models and multivariate classifiers to detect anomalous network conditions. [C Manikopoulos, 2002]. Many anomaly detection techniques have been developed and evaluated in the last several years but reducing false alarms is still a challenging task in anomaly detection.[Eskin et al.] analyzed three different algorithms for unsupervised anomaly detection: cluster based estimation, k-nearest neighbor, and one class SVM (Support Vector Machine). Other supervised anomaly detection techniques include ADAM (Audit Data Analysis and Mining), neural networks, and SVM [Q. A. Tran et al.]. ADAM is a well-known on-line network based IDS. It can detect known as well as unknown attacks. It builds the profile of normal behavior from attack-free training data and represents the profile as a set of association rules. It detects suspicious connections according to the profile.

## VI. ANOMALY DETECTION PROCEDURE

The nemesis of anomaly-based detection has been the false positive. A detection system cannot be perfect (even if it uses a human expert). It produces false positive (it thinks it has detected a malicious event, which in fact is legitimate) and has false negative (it fails to detect actual malicious event).

A. Proposed Architecture of Anomaly Detection

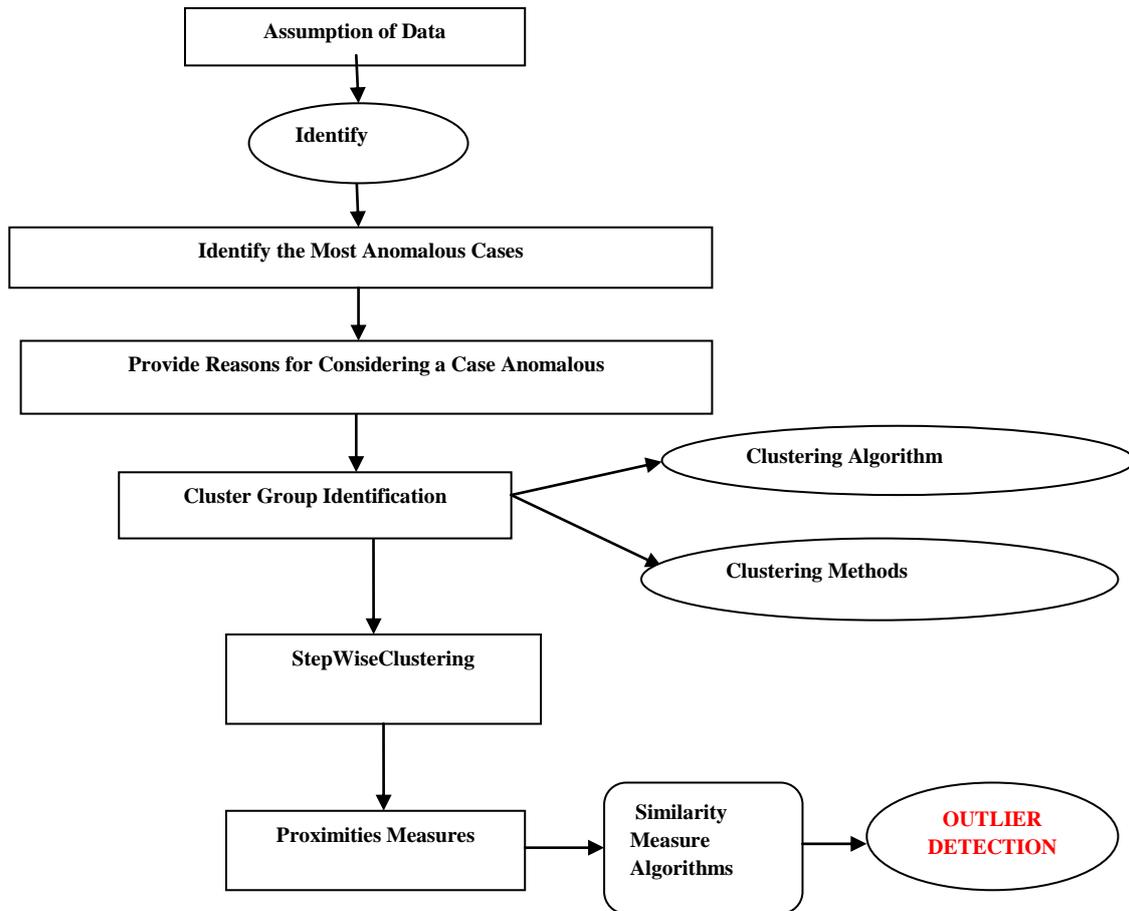


Figure 1. Proposed Architecture of Anomaly Detection

B. Algorithm

**Step 1:**

**Assumption of Data:**

This procedure can be used for both continuous and the categorical variables. Distinct observation of values represented in rows and distinct variable represented in columns each respectively based on cluster. An identification of attack are stored as a variable in the data auditing file for the output, but not used for the purpose of analysis. The anomaly detection method can be applied to the new set of data for testing, it be the same as the training data elements. The detection procedure assumes that all variables are not constants and independent. Each case has continuous variable is assumed to have a normal distribution, and each categorical variable have a multinomial distribution.

**Step 2:**

**Identify:**

Each and every unusual case of attack now has a group deviation index and anomaly index and a set of variable deviation index and distance variable measure. The purpose of this step of procedure is to rank the likely anomalous cases and to provide the reasons to suspect them of being the anomalous attack.

1. **Identify the Most Anomalous Cases.** Sort the cases in descending order on the values of the anomaly index.

2. **Provide Reasons for Considering a Case Anomalous.** For each anomalous case, sort the variables by their corresponding Variable Deviation Index values in descending order.

### Step 3: Cluster Group Identification:

The *StepWiseClustering* Algorithm is used to create the clustering model for the processed input variables.

- **Clustering Algorithm:** CLUSTER produces hierarchical clusters of items based on distance measures of dissimilarity or similarity.
- Begin with N clusters each containing one case. Denote the clusters 1 through N.
- Find the most similar pair of clusters p and q ( $p > q$ ) denote this similarity. If a dissimilarity measure is used, large values indicate dissimilarity. If a similarity measure is used, small values indicate dissimilarity.
- Reduce the number of clusters by one through merger of clusters p and q. Label the new cluster t and update similarity matrix (by the method specified) to reflect revised similarities or dissimilarities between cluster t and all other clusters. Delete the row and column of S corresponding to cluster p.
- Perform the previous two steps until all entities are in one cluster.
- **Clustering Methods:** The cluster method defines the rules for cluster formation. For example, when calculating the distance between two clusters, you can use the pair of nearest objects between clusters or the pair of furthest objects, or a compromise between these methods.

Example:

$N_1$  and  $N_2$  are regions of normal behaviour

Point  $o_1$  and  $o_2$  are anomalies

Points in region  $O_3$  are anomalies

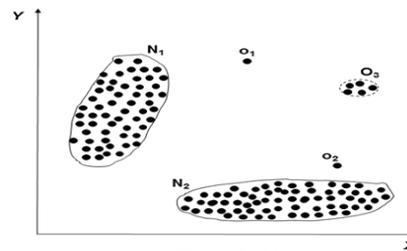


Figure 2. Clusters

### Step 4: StepWiseClustering:

The *StepWiseClustering* algorithm is a measurable cluster analysis algorithm designed for the large set of data analysis. It can use and handle both the continuous variable and categorical variables. It takes one pass method of data. It has two step processes,

1. Preprocessed cluster the cases into many small sub-clusters.
2. Cluster the small sub-cluster cases into the desired number of clusters

### Step5: Proximities Measures:

Proximities measure defines the formula for calculating distance.

For example, the Euclidean distance measure calculates the distance as a “straight line” between two clusters. The distance between two items, x and y, is the square root of the sum of the squared differences between the values for the items.

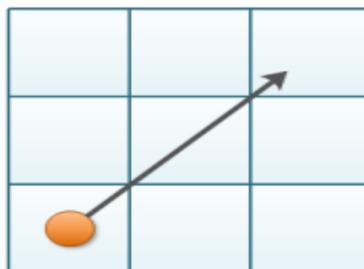


Figure 3. Euclidean Distances

C. Program for distance calculation by Euclidean distance formula between the cluster points

```
#include<iostream>
```

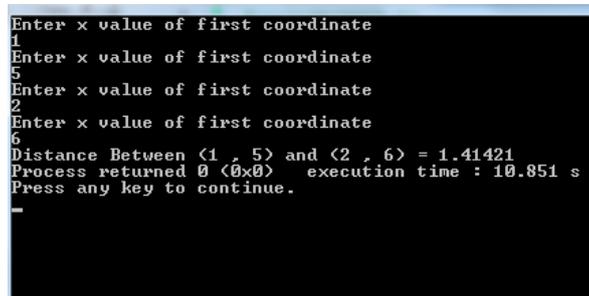
```
#include<conio.h>
```

```

#include<math.h>
void main()
{
    double x1, y1, x2, y2,x,y;
    double dist;
    cout<<"Enter x value of first coordinate "<<endl;
    cin>>x1;
    cout<<"Enter y value of first coordinate "<<endl;
    cin>>y1;
    cout<<"Enter x value of second coordinate "<<endl;
    cin>>x2;
    cout<<"Enter y value of second coordinate "<<endl;
    cin>>y2;
    x = x1 - x2;
    y = y1 - y2;
    dist = pow(x,2)+pow(y,2);    //calculating distance by Euclidean formula
    dist = sqrt(dist);
    cout<<"Distance Between ("<<x1<<" , "<<y1<<") and ("<<x2<<" , "<<y2<<") = "<<dist;
}

```

Output:



```

Enter x value of first coordinate
1
Enter x value of first coordinate
5
Enter x value of first coordinate
2
Enter x value of first coordinate
6
Distance Between (1 , 5) and (2 , 6) = 1.41421
Process returned 0 (0x0)   execution time : 10.851 s
Press any key to continue.

```

Figure 4. Distance Calculation by Euclidean Distance Formula

### Proximities Algorithms:

**Kulczynski Similarity Measure 2:** This yields the average conditional probability that a characteristic is present in one item given that the characteristic is present in the other item. The measure is an average over both items acting as predictors. It has a range of 0 to 1 [11].

$$K2(x, y) = \frac{a/(a+b) + a/(a+c)}{2} \quad (1)$$

**Sokal and Sneath Similarity Measure 4:** This yields the conditional probability that a characteristic of one item is in the same state (present or absent) as the characteristic of the other item. The measure is an average over both items acting as predictors. It has a range of 0 to 1.

$$SS4(x, y) = \frac{a/(a+b) + a/(a+c) + d/(b+d) + d/(c+d)}{4} \quad (2)$$

**Hamann Similarity Measure:** This measure gives the probability that a characteristic has the same state in both items (present in both or absent from both) minus the probability that a characteristic has different states in the two items (present in one and absent from the other). HAMANN has a range of -1 to +1 and is monotonically related to SM, SS1, and RT.

$$HAMANN(x, y) = \frac{(a+d) - (b+c)}{a+b+c+d} \quad (3)$$

**Step 6: Similarity Measure Algorithm:**

Hierarchical agglomerative clustering:

- We start with every data point in a separate cluster.
- We keep merging the most similar pairs of data points/clusters until we have one big cluster left.
- This is called a bottom-up or agglomerative method.

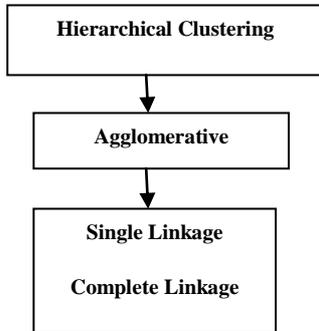


Figure 5. HA Methods

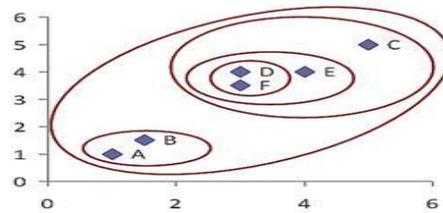


Figure 6. Sample Clustering

A *dendrogram* is a tree diagram it is frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering [Wikipedia]. The Strategies for hierarchical clustering generally fall into two types. Efficient agglomerative methods with complexity  $O(n^2)$  are SLINK and CLINK.

The linkage criterion determines the distance between sets of observations as a function of the pair wise distances between observations.

**Step 7: Outliers Detections:**

- *Detection rate* - ratio between the number of correctly detected anomalies and the total number of anomalies.
- *False alarm (false positive) rate* - ratio between the number of data records from normal class that are misclassified as anomalies and the total number of data records from normal class.
- *ROC Curve* is a trade-off between detection rate and false alarm rate.
- *Area under the ROC curve (AUC)* is computed using a trapezoid rules.

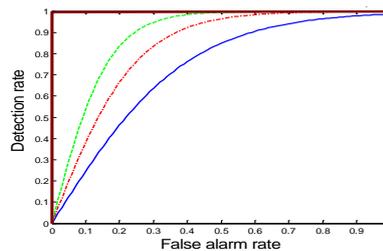


Figure 7. ROC Curve

**VII. CONCLUSION**

Anomaly detection refers to the problem of finding unusual patterns in data. These patterns are often known as anomalies, outliers, exceptions. In anomaly detection using the clustering and proximities measure is an effective technique of anomalies in networks. It initially clusters the normal data using the Stepwise clustering algorithm and clustering methods. Using the proximities measuring technique calculates the distance using the Euclidean distance measure and also the reference point from each cluster and builds profiles for each cluster. It calculates the score for each point with respect to the reference point. A dendrogram is a tree diagram it is used to illustrate the arrangement of the clusters produced by hierarchical clustering. Performance of our technique has been evaluated using intrusion datasets. For further extending the paper use modified classification model to learn the normal behaviour and then detect any deviations from normal behaviour as anomalous.

## REFERENCES

- [1] Benoît Morel, "Anomaly Based Intrusion Detection and Artificial Intelligence", Carnegie Mellon University, United States, pp. 21-38, 2011.
- [2] Fengmin Gong, "Deciphering Detection Techniques: Part II Anomaly-Based Intrusion Detection", International Journal for Advances in Computer Science, vol. 4, pp. 436-444, 2003.
- [3] Saira Beg, "Feasibility of Intrusion Detection System with High Performance Computing: A Survey", International Journal for Advances in Computer Science, vol. 3, pp. 406-414, 2010.
- [4] Vera Marinova-Boncheva, "A Short Survey of Intrusion Detection Systems", Problems of Engineering Cybernetics and Robotics, vol. 58, pp. 23-30, 2007.
- [5] Kenneth D. Jarman, "Integrating Correlated Bayesian Networks Using Maximum Entropy", Applied Mathematical Sciences, vol. 5, pp. 2361 – 2371, 2011.
- [6] Jing Xu, Christian R, "Shelton Intrusion Detection using Continuous Time Bayesian Networks", Journal of Artificial Intelligence Research, vol. 39, pp. 745–774, 2010.
- [7] Pablo G. "Bringas and Igor Santos Bayesian Networks for Network Intrusion Detection", University of Deusto, pp. 229-244, 2010.
- [8] M. Mehdi, "A Bayesian Networks in Intrusion Detection Systems Electronics Department", University of Blida, 2007.
- [9] P. Garcia, "Anomaly-based network intrusion detection: Techniques, systems and challenges", International Journal of Advanced Sciences and Technology, vol. 28, pp. 18-28, 2009.
- [10] D. Djenouri, L. Khelladi, and N. Badache, "A survey of security issues in mobile ad hoc and senso networks," IEEE Communications Surveys & Tutorials, vol. 7, pp. 2-28, 2005.
- [11] M. S. Islam and S. A. Rahman, "Anomaly intrusion detection system in wireless sensor networks: security threats and existing approaches," International Journal of Advanced Sciences and Technology, vol. 36, pp. 245-256, 2011.