# Anomaly Detection using Outlier Detection Methods: A Survey.

**Sunil Kumar Rajwar,**

*Assistant Professor, University Department of Computer Applications, Vinoba Bhave University, Jharkhand, India.*

**Dr. I Mukherjee**

*Assistant Professor, Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi ,Jharkhand , India*

**Dr. Pankaj Kumar Manjhi**

*Assistant Professor, University Department of Mathematics, Vinoba Bhave University, Jharkhand, India*

**ABSTRACT**

*The importance of Outlier Detection is due to the fact that Outlier in data translate to significant actionable information in wide variety of application domains [3]. Computer network intrusion detection system is developed to identify a deviation to a known behaviour through networking monitoring activities. An Intrusion Detection System (IDS) is usually used to enhance the network security of enterprises by monitoring and analysing network data packets. Outlier based network intrusion detection techniques are a valuable technology to protect target systems and network against malicious activities [10].This paper present a comprehensive survey of well known techniques for Outlier Detection.*

***KEYWORDS: Outlier,Intrusion Detection Sysytem,Network Security***.

## 1. INTRODUCTION

Outlier detection also known as anomaly detection is an important research problem in data mining that aims to find objects that are considerably dissimilar, exceptional and inconsistent with respect to the majority data in an input database [1]. Outlier is defined as an observation (or subset of observation) which appears to be inconsistent with the remainder of that set of data [2]. Outlier detection have been widely used for various applications such as Intrusion Detection System , Cyber Security , Geographic Information systems, Fraud Detection, Web Data Analysis, Surveillance, Economics and Time Series Data Analysis.

Outliers are different from noisy data. Noise is a random error or variance in a measured variable. Noise is not interesting in data analysis including outlier detection [1]. Outlier will arise due to natural variability of data set, measurement error as well as recording error done by the users and execution error [5]. This paper presents a comprehensive survey on major outlier detection methods. We will cover major categories of outlier detection methods and their respective advantages and disadvantages.

1.1 Outlier Detection

Anomalies or outliers are pattern in data that do not conform to a well defined notion of normal behaviour[3].Figure 1. Illustrates outlier in two dimensional data sets. The data has various normal regions(N1,N2,N3),most observations lie in these regions. Points or data that are sufficiently away from the regions are outliers(O1,O2,O3).

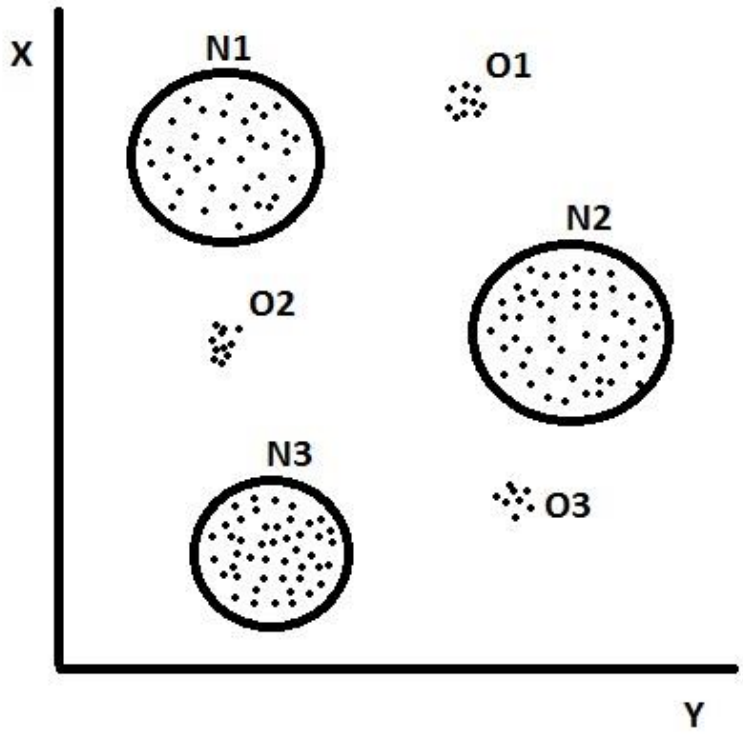


Fig 1. Anomaly or Outlier in a data set.

1.2 Type of outliers:

In general outliers can be classified in three categories: Global outliers, Contextual Outliers and Collective Outliers [1].

Global Outliers:

In a given data set, a data object is a global outlier if it deviates significantly from the rest of the data set. it is also called point outlier.

Contextual outlier:

In a given data set, a data object is a contextual outlier if it deviates significantly with respect to a specific context of the object. Contextual outliers are also known as conditional outliers because they are conditional on the selected context.

Collective outliers:

In a given data set , a subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set.

1.3 General Framework for Network Outlier Detection :

An *ANIDS* is an anomaly based network intrusion detection system [11]. Although different A-NIDS Approaches developed so far [10][12][13], in general terms all of them consist of the following modules or stages.

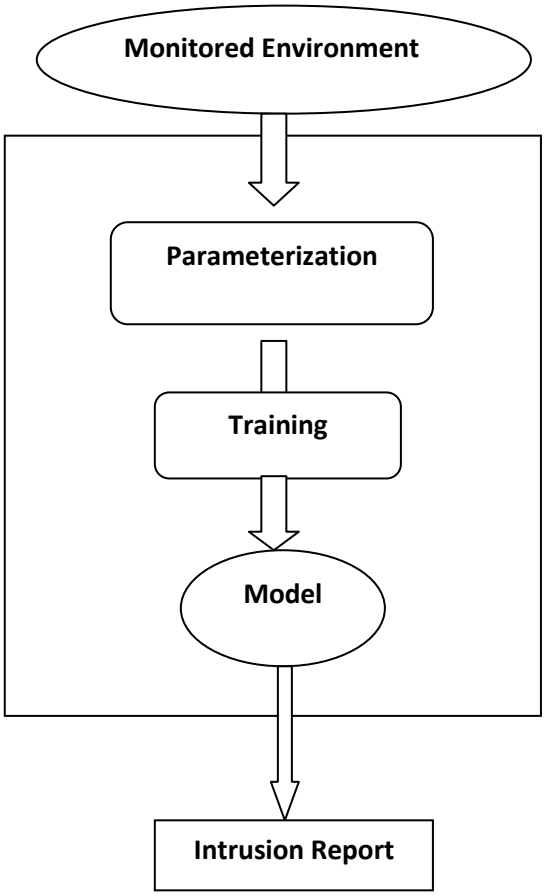


Fig 2: General A-NIDS Architecture.

## 2. OUTLIER DETECTION TECHNIQUES.

In this paper we are describing the different approaches and methods used for outlier detection along with their advantages and disadvantages. Basically there are two approaches are used in outlier detection : Supervised and Unsupervised. Supervised outlier detection techniques assume the availability of a training data set which has labelled instances for the normal as well as the outlier class. In such techniques, predictive models are built for both normal and outlier classes. Any unseen data instance is compared against the two models to determine which class it belongs to. An unsupervised outlier detection technique makes no assumption about the availability of labelled training data. Thus, these techniques are more widely applicable. The techniques in this class make other assumptions about the data [11].

This figure explains different methods used for outlier detection according the data set and based on different characteristics. this figure generalizes the basic methods used for outlier detection on different data objects[8][9].

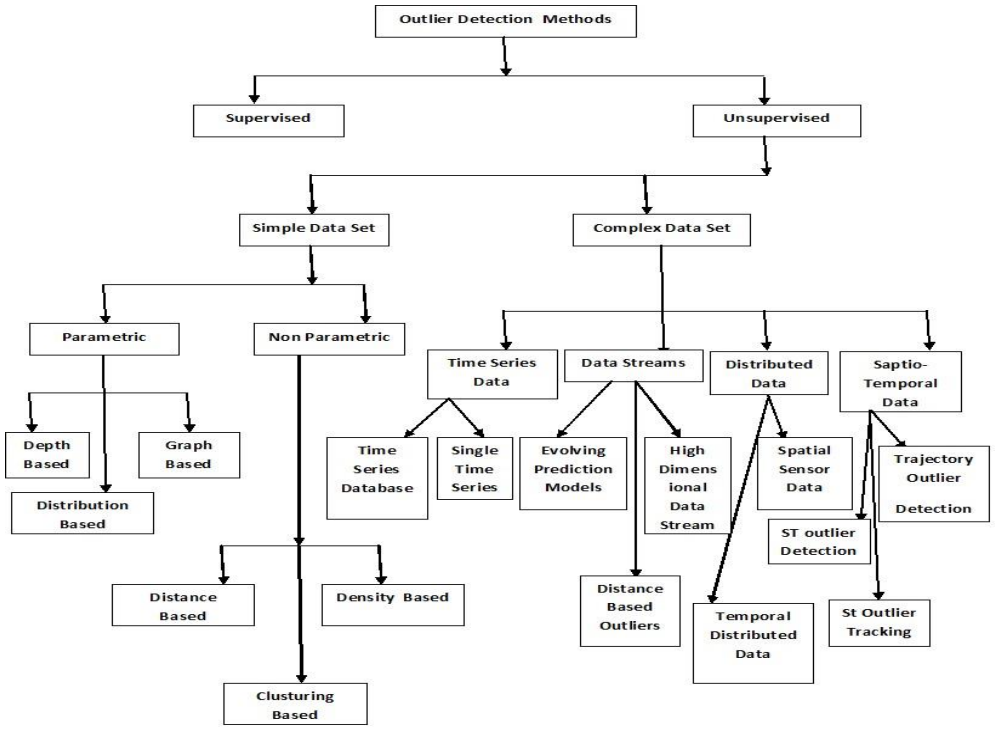


Fig 3: Structure of Outlier Detection Methods

2.1 Classification based Outlier detection:

Classification is the process of assigning a class label to unclassified object based on a set of defined features. A classifier should first get that knowledge by learning the representation of classes using a given set of pre-classified samples. A classifier would act as predictor for unclassified objects or descriptor for classified objects. Different approaches are used such as Decision Trees, Rule Based Approaches. Bayesian Classifiers, Neural Networks, Genetic Classifiers, Support Vector Machines.[15][16][17].

2.1.1 Decision Tree

Decision Trees are Trees that classify instances by sorting them based on feature values [18].Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume . Instances are classified starting at the root node and sorted based on their feature values. DTs are very powerful tools because they are fast and give reasonable performance. Different DTs applied are Best-First Tree, Naive-Bayes Tree and Random Forrest Trees [14].

2.1.2 Multi-layer perceptrons.

Perceptrons can only classify linearly separable sets of instances. If the instances are not linearly separable learning will never reach a point where all instances are classifieds properly. A Multi-Layer Neural network consists of large number of units (neurons ) joined together in a pattern of connections units in a NN are usually segregated into three classes: Input units which receive information to be processed. Output units where result of the processing are found and units in between known as hidden units.ANN depends on three fundamental aspects, input and activation functions of unit, network architecture and the weight of each input connection. The well known and widely used learning algorithms used in ANN is Back Propagation algorithm [18].

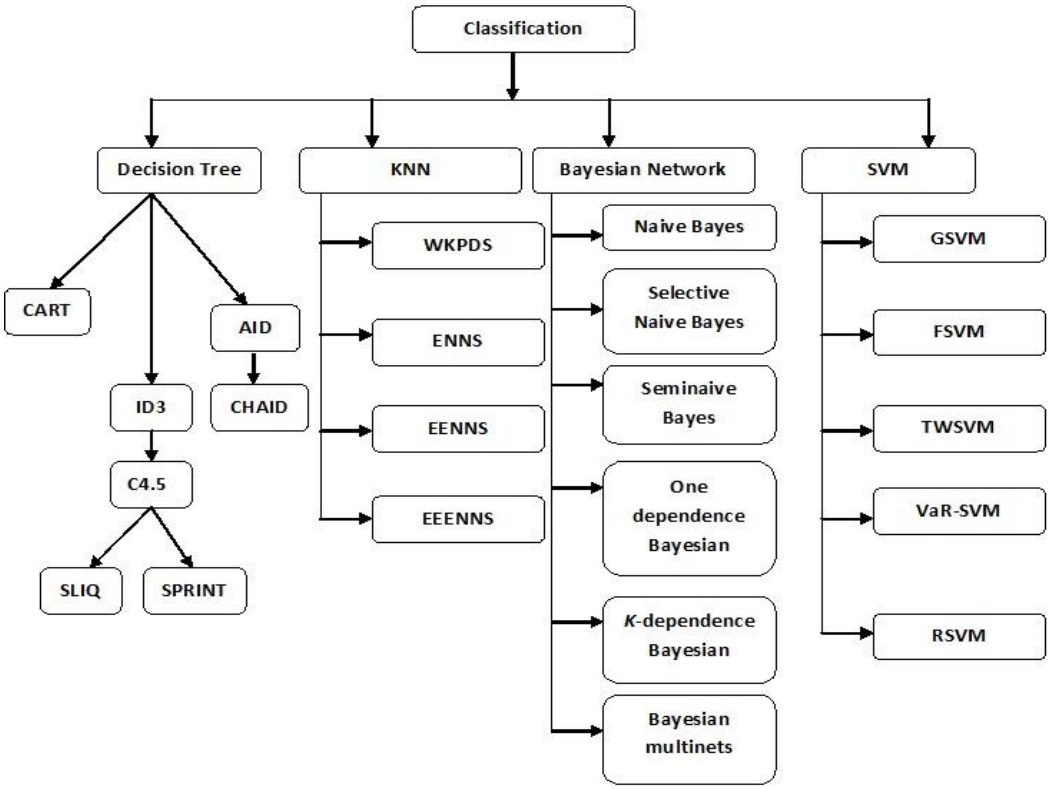


Fig 4: Structure of Classification for Outlier Detection [14].

2.2 Clustering based Outlier Detection

Clustering is a technique for finding patterns in unlabelled data with many dimensions equivalent to the number of attributes [19].Clustering can be defined as a division of data into groups of similar objects. Each cluster consists of object that are similar to one another and dissimilar to objects in other groups.[20][13]. There are different approaches are used for Clustering which are as follows:

2.2.1 Hierarchical Clustering:

Hierarchical clustering method combines data objects into subgroups; those subgroups merge into high level groups and so forth and form a hierarchy tree. CURE (Clustering using Representatives) and SVD (Singular value Decomposition) are different research in Hierarchical Clustering.

2.2.2 Partitioning Clustering Methods:

The Clustering Methods using partitioning perform clustering by partitioning the data set into specific number of clusters. The partitioning clustering methods are PAM, CLARA, k-means and CLARANS [21].

2.2.3 Density based Clustering methods:

The density based clustering methods consider normal clusters as dense regions of the objects in the data space. The density based clustering algorithms are DBSCAN and DENCLUE.

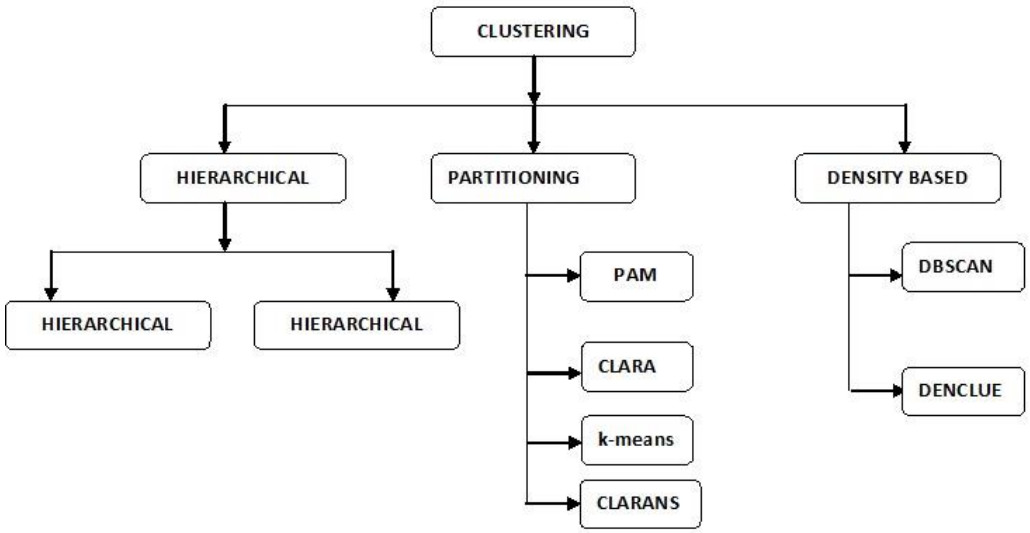


Fig 5: Structure of Clustering methods for Outlier detection.

2.3 Hybrid Methods for Outlier Detection.

The monitoring capability of current Network Intrusion Detection Systems(NIDS) can be improved by applying hybrid approach that consists of both outlier detection as well as signature detection methods[19].Combinations of k means and ID3 was proposed for classification of anomalous and normal activities in computer Address Resolution Protocol(ARP) traffic and 98 percent of accuracy was achieved[13][23].

EMERALD is a hierarchal intrusion detection system that monitors systems at a variety of levels. The most defining features of EMERALD is its ability to analyse attacks such as Internet worms, DDoS attacks [19].The major challenge to build an operational hybrid intrusion detection system is getting these different technologies to interoperate and efficiently.

## CONCLUSION:

Through this paper various outlier detection techniques for network anomaly detection has been described for the last 20 years. This paper helps the researcher to gain different new technological developments in network intrusion detection systems and their different detection methods. The challenges that lie ahead for the next generation of intrusion detection systems are many. Noise in the audited data constantly changing traffic profiles and the large amount of network traffic make it difficult to build normal traffic profile for outlier detection in intrusion detection for network. Although there are so many existing work in this area but more research is needed for practical solutions.

## REFERENCES

1. J.Han and M.Kamber,2000,Data Mining:Concepts and Techniques,Morgan Kaufman Publishers.
2. Barnett V, Lewis T, 1994, Outliers in Statistical Data,New York NY: John Wiley & Sons:.
3. Varun Chandola,A Banerjee,2009,Vipin Kumar, Anmoly Detection:A Survey. ACM Computing Surveys.
4. M J Liao, C-H Richard Lin, Y-C Lin, K-Y Tung,2013,Intrusion Detection System: A Comprehensive Review, Journal of Network and Computer Applications,Elsevier.
5. Huang,H.Y,Lin J.X, Chen C.C and Fan M.H,2006, Review of Outlier Detection,In: Application Research of Computer.
6. M.Markov and S Singh ,2003,Novelty detection review part1,Signal Processing.
7. M.Markov and S Singh ,2003,Novelty detection review part2,Signal Processing.
8. Charu. C Agarwal,Manish Gupta, Jing Gao and Jawei Han,2014,Outlier Detection for Temporal Data:A Survey,IEEE Transactions on Knowledge and Data Engineering.
9. V.Hango,R Subramanian, V. Vasudevan, 2012,A Five step Procedure for Outlier Analysis in Data Mining.European Journal of Scientific Research.

10. P. Garcia-Teodoro, J Diaz-Verdigo , G macia-Fernandez,E. Vazquez, 2008,Anomaly based Network Intrusion Detection :Techniques,System and Challenges. Computer & Security.
11. P. Gogoi, D.K Bhattacharya, Borah and J.K Kalita.2011 A Survey of Outlier Detection Methods in Network Anomaly Identification,The Computer Journal.
12. Estevez-Tapisdeo JM, Garcia-Teodoro P, Diaz-verdejo J E,2004,Anomaly Detection Methods in Wired Networks: A Survey and Taxanomy,Computer Networks.
13. Shikha Agarwal, Jitendra Agarwal ,2015,Survey on Anomaly Detection using Data Mining Techniques,Procedia Computer Science.
14. Feng Chen, Pan Deng ,Jiafu Wan ,D Zhang, Athanasios V. Vasilakos and Xiaohui Rong,2015,Data Mining for Internet of Things: Litereature Review and Challenges,International Journal of Distributed Sensor Networks,Hindawi Publishing Corp.
15. Amira sayed A. Aziz, Sanaa EL-ola Hanafi, Aboul Ella Hassanien, 2016,Comparison of Classification Techniques applied for Network intrusion detection and Classification,Journal of Logic, Elsevier.
16. S. Ganapathy, K. Kulothungan, S. MuthurajKumar,M.Vijayalakshmi, P.Yogesh, A.Kahnan,2013, Intelligent Feature Selection and Classification Techniques for Intrusion Detection in Network: A Survey, EURASIP J-Wirel Comm. Net.
17. C.So-In, N Mongkonchai, P.Aimtongkham, N. Wijitsopon, K.Rajirakul, 2014, An Evaluation of Data Mining Classification Models for Network Intrusion Detection ,DICTAP,IEEE.
18. S.B Kotsiantis,2007,Supervised Machine Learning:A Review of Classification Techniques,Informatica.
19. Animesh Patcha,Jung-Min Park,2007, An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends, Computer Networks,Elsevier.
20. Berkhin P,2006, A Survey of Clustering Data Mining Techniques:Grouping Multidimensional Data,Springer Berlin Heidelbeg.
21. Ji Zhang,2013, Advancements of Outlier Detection : A survey, ICST Transactions on Scalable Information Systems.
22. Yasmi Y., Mozaaffari S.P,2010, A novel Unsupervised Classification Approach For Network Anomaly Detection by k-means Clustering and ID3 Decision Tree Methods,The Journal of Supercomputing.