IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. III (Jan – Feb. 2015), PP 25-29
www.iosrjournals.org

Performance Analysis of Different Clustering Algorithm

[1]Naresh Mathur, [2]Manish Tiwari, [3]Sarika Khandelwal
[1]M.Tech Scholar, [2]Assistant Professor, [3]Associate Professor
[1,2&3]Department of Computer Science Engineering, Geetanjali Institute of Technical Studies, Udaipur.

Abstract: *Clustering is the process of grouping objects into clusters such that the objects from the same clusters are similar and objects from different clusters are dissimilar. The relationship is often expressed as similarity or dissimilarity measurement and is calculated through distance function. some of the outlier detection techniques are distance based outlier detection distribution based outlier detection density based outlier detection and depth based outlier detection The goal of this paper is the detection of outliers with high accuracy and time efficiency. The methodology discussed here is able to save a large amount of time by selecting a small subset of suspicious transactions for manual inspection which includes most of the erroneous transactions.*

Keywords: *PAM.CLARA, CLARANS, ECLARANS*

I. Introduction

Data mining is the method of extracting patterns from data. It can be used to uncover patterns in data but is often carried out only on sample of data. analysis is a tool for exploring the structure of data. Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster Clustering is the process of grouping objects into clusters such that the objects from the same clusters are similar and objects from different clusters are dissimilar. The relationship is often expressed as similarity or dissimilarity measurement and is calculated through distance function. Clustering is useful technique for the discovery of data distribution and patterns in the underlying data. are more similar to each other than they are to a pattern belonging to a different cluster.

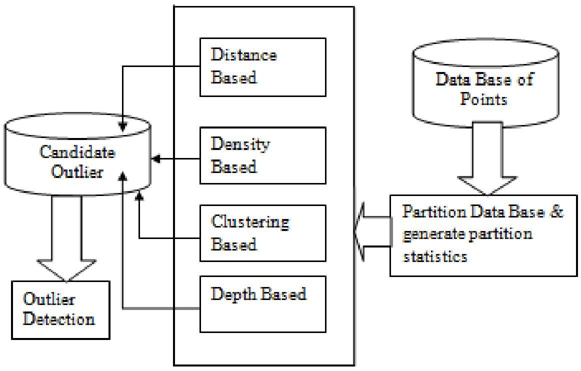


Fig 1: Outlier Detection Module

Outliers detection is an outstanding data mining task, referred to as outlier mining. Outliers are objects that do not comply with the general behavior of the data. By definition, outliers are rare occurrences and hence represent a small portion of the data.

II. Categorization Of Clustering Techniques

According to Data Mining concepts and Techniques by Jiawai Han and Micheline Kamber clustering algorithm partition the dataset into optimal number of clusters.

They introduce a new cluster validation criterion based on the geometric property of data partition of the dataset in order to find the proper number of clusters. The algorithm works in two stages. The first stage of the algorithm creates optimal number of clusters , where as the second stage of the algorithm detect outliers.

2.1 Cluster Algorithms:

Algorithms which are being used for outlier detection are-

- PAM(Partitioning around Medoids)
- CLARA(Clustering large applications)

- CLARANS(Clustering large applications by randomized search)
- ECLARANS(Enhanced Clarans)

2.1.1 PAM (Partitioning Around Medoids) –

PAM (Partitioning Around Medoids) was developed by Kaufman and Rousseeuw. To find k clusters, PAM's approach is to determine a representative object for each cluster. This representative object, called a medoid, is mean to be the most centrally located object within the cluster. Once the Medoids have been selected, each non selected object is grouped with the medoid to which it is the most similar.

Procedure-

1. Input the dataset D
2. Randomly select k objects from the dataset D
3. Calculate the Total cost T for each pair of selected Si and non selected object Sh
4. For each pair if T si < 0, then it is replaced Sh
5. Then find similar medoid for each non-selected object 6. Repeat steps 2, 3 and 4, until find the Medoids.

2.1.2 CLARA (Clustering large applications)-

Designed by Kaufman and Rousseeuw to handle large datasets, CLARA (Clustering large Applications) relies on sampling. Instead of finding representative objects for the entire data set, CLARA draws a sample of the data set, applies PAM on the sample, and finds the Medoids of the sample. The point is that, if the sample is drawn in a sufficiently random way, the Medoids of the sample would approximate the Medoids of the entire data set. To come up with better approximations, CLARA draws multiple samples and gives the best clustering as the output. Here, for accuracy, the quality of a clustering is measured based on the average dissimilarity of all objects in the entire data set, and not only of those objects in the samples.

CLARA Procedure-

1. Input the dataset D
2. Repeat n times
3. Draw sample S randomly from D
4. Call PAM from S to get Medoids M.
5. Classify the entire dataset D to Cost1.....cost k
6. Calculate the average dissimilarity from the obtained clusters

Complementary to PAM, CLARA performs satisfactorily for large data sets (e.g., 1,000 objects in 10 clusters).

2.1.3 CLARANS (A clustering algorithm based on randomized search)

It gives higher quality clusterings than CLARA, and CLARANS requires a very small number of searches. We now present the details of Algorithm CLARANS.

Procedure of CLARANS-

1. Input parameters num local and max neighbour. Initialize i to 1, and min cost to a large number.
2. Set current to an arbitrary node in n:k.
3. Set j to 1.
4. Consider a random neighbour S of current, and based on 5, calculate the cost differential of the two nodes.
5. If S has a lower cost, set current to S, and go to Step 3.
6. Otherwise, increment j by 1. If j max neighbour, go to Step 4.
7. Otherwise, when j > max neighbour, compare the cost of current with min cost. If the former is less than min cost, set min cost to the cost of current and set best node to current.
8. Increment i by 1. If i > num local, output best node and halt. Otherwise, go to Step 2.

Steps 3 to 6 above search for nodes with progressively lower costs. But, if the current node has already been compared with the maximum number of the neighbors of the node (specified by max neighbor) and is still of the lowest cost, the current node is declared to be a “local” minimum. Then, in Step 7, the cost of this local minimum is compared with the lowest cost obtained so far. The lower of the two costs above is stored in min cost. Algorithm CLARANS then repeats to search for other local minima, until num local of them have been found.

As shown above, CLARANS has two parameters: the maximum number of neighbours examined (max neighbour) and the number of local minima obtained (num local). The higher the value of max neighbour, the closer is CLARANS to PAM, and the longer is each search of a local minima. But, the quality of such a local minima is higher and fewer local minima need to be obtained.

III. Proposed Work

The procedure followed by partitioning algorithms can be stated as follows: “Given n objects, these methods construct k partitions of the data, by assigning objects to groups, with each partition representing a cluster. Generally, each cluster must contain at least one object; and each object may belong to one and only one cluster, although this can be relaxed”. The present study analyzes the use of PAM , CLARA, CLARANS and ECLARNS.

ENHANCED CLARANS (ECLARANS): This method is different from PAM, CLARA AND CLARANS. Thus method is produced to improve the accuracy of outliers. ECLARANS is a partitioning algorithm which is an improvement of CLARANS to form clusters with selecting proper nodes instead of selecting as random searching operations. The algorithm is similar to CLARANS but these selected nodes reduce the number of iterations of CLARANS ECLARANS Procedure. The Previous research established ECLARANS as an effective algorithm for outlier detection but till now it doesn’t have better time complexity thus by this research work we can also achieve this.

The algorithm is-

1. Input parameters num local and max neighbour. Initialize i to 1, and min cost to a large number.
2. Calculating distance between each data points
3. Choose n maximum distance data points
4. Set current to an arbitrary node in n: k
5. Set j to 1.
6. Consider a random neighbour S of current, and based on 6, calculate the cost differential of the two nodes.
7. If S has a lower cost, set current to S, and go to Step 5.
8. Otherwise, increment j by 1. If j max neighbour, go to Step 6.
9. Otherwise, when j > max neighbour, compare the cost of current with min cost. If the former is less than min cost, set min cost to the cost of current and set best node to current.
10. Increment i by 1. If i > num local, output best node and halt. Otherwise, go to Step 4.

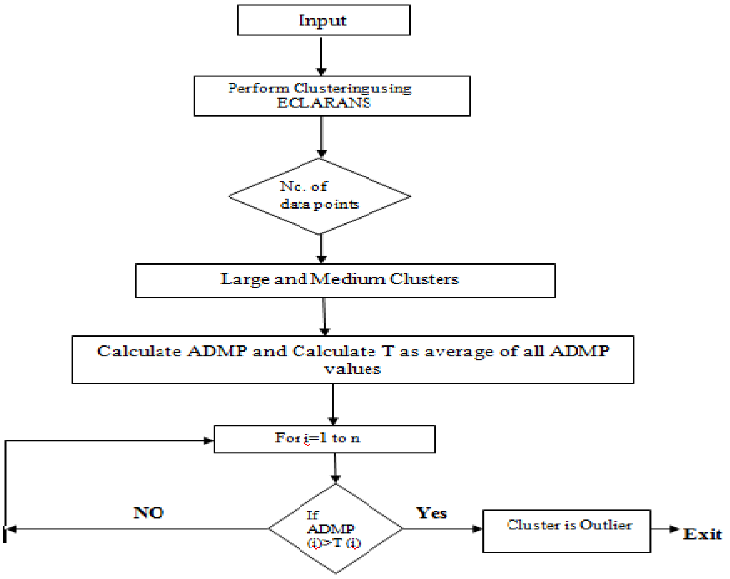

Input
Perform Clusteringusing ECLARANS
No. of data points
Large and Medium Clusters
Calculate ADMP and Calculate T as average of all ADMP values
For i=1 to n
NO
If ADMP (i)>T (i)
Yes
Cluster is Outlier
Exit

Fig 2: Flowchart of ECLARANS algorithm

3.1 Proposed Methodology

In modified ECLARANS the approach of selecting nodes have been changed rather than selecting random nodes after calculating the maximum cost between nodes we have chosen that points which are causing maximum cost.

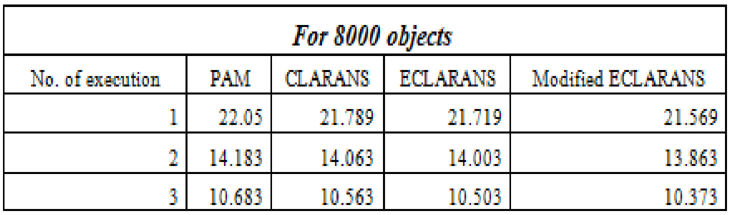
3.2 Modified Algorithm

1. Input parameters num local and max neighbour. Initialize i to 1, and min cost to a large number.
2. Calculating distance between each data points for calculation select those points which has not been visited.
3. Select the maximum distance data points.
4. Set current to that node which is having highest distance if it is not been visited.
5. Set j to 1.

6. Consider a random neighbour S of current, and based on 6, calculate the cost differential Between two nodes.
7. If S has a lower cost, set current to S, and go to Step 5.
8. Otherwise, increment j by 1. If j max neighbour, go to Step 6.
9. Otherwise, when j > max neighbour, compare the cost of current with min cost. If the former is less than min cost, set min cost to the cost of current and set best node to current.
10. Increment i by 1. If i > num local, output best node and halt. Otherwise, go to Step4.

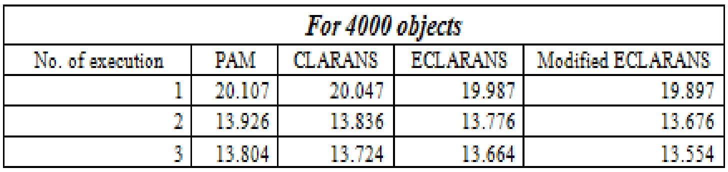
IV. Results

Proposed algorithm has been implemented using programming language Java and WHO dataset. The java has been used using Netbeans7.3.1 which provides easy to implement graphical user interface for the proposed system. Implemented software has been run using various lengths of data. Time required for various executions has been recorded for different steps of the proposed work and results have been drawn. After running various time the table of execution time (in seconds) are-



For 8000 objects				
No. of execution	PAM	CLARANS	ECLARANS	Modified ECLARANS
1	22.05	21.789	21.719	21.569
2	14.183	14.063	14.003	13.863
3	10.683	10.563	10.503	10.373

Table5.1 Execution Time for 8000 Objects



For 4000 objects				
No. of execution	PAM	CLARANS	ECLARANS	Modified ECLARANS
1	20.107	20.047	19.987	19.897
2	13.926	13.836	13.776	13.676
3	13.804	13.724	13.664	13.554

Table5.1 Execution Time for 4000 Objects

By the no. of execution analysis chart has been created which is a graphical presentation of results used for comparison with other clustering algorithms

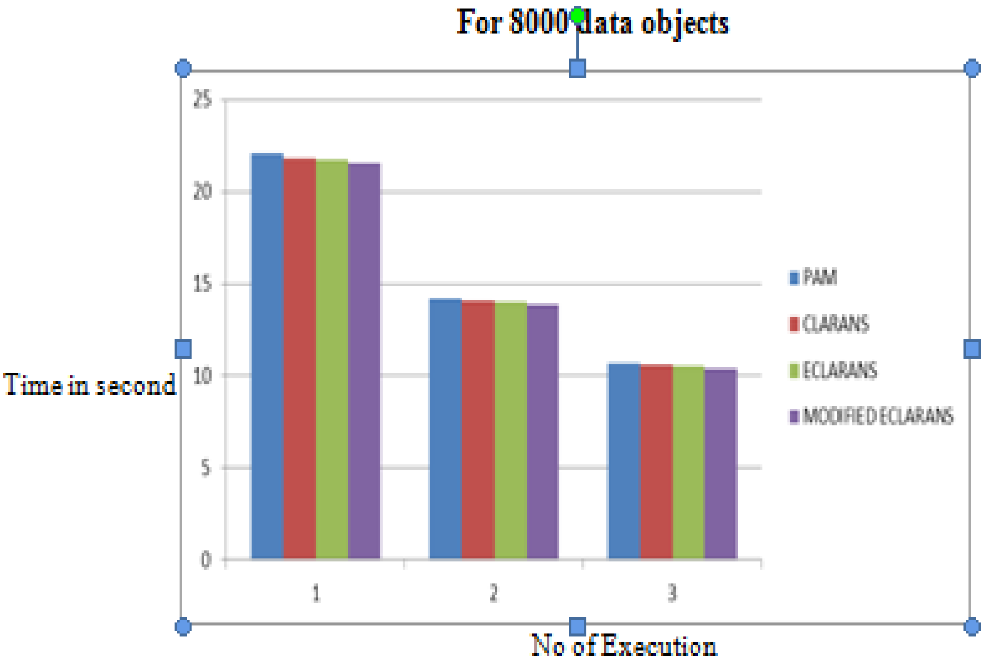


Fig 3: comparison of Different clustering algorithms for 8000 data objects

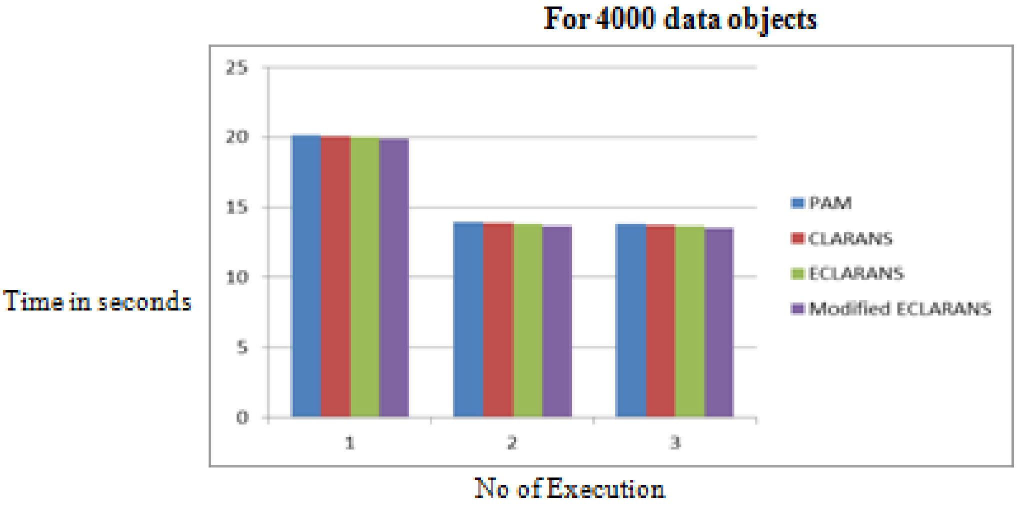


Fig 4: comparison of Different clustering algorithms for 8000 data objects

V. Conclusion

Modified ECLARANS has been found more accurate and time efficient. There are large number of Partition based outlier detection technique are available. They can be used to solve all problems. But all algorithms are designed under certain assumption and different algorithm is used under different condition. Such as k-mean is used to handle spherical shaped cluster we cannot used to find arbitrary shaped cluster. The main aim of this clustering algorithm is , outlier detection with improved time efficiency and outlier detection accuracy. Additionally, the efficiency and effectiveness of a novel outlier detection algorithm can be defined as to handle large volume of data as well as high-dimensional features with acceptable time and storage, to detect outliers in different density regions, to show good data visualization and provide users with results that can simplify further analysis.

References

[1]. A. Mira, D.K. Bhattacharyya, S. Saharia,” RODHA: Robust Outlier Detection using Hybrid Approach”, American Journal of Intelligent Systems, volume 2, pp 129-140, 2012

[2]. Al-Zoubi M. “An Effective Clustering-Based Approach for Outlier Detection”(2009)

[3]. A K Jain,M N Murthy. “Data Clustering A Review” ACN Computing Surveys Vol 31,No3.September 1999.

[4]. D Moh, Belal Al-Zoubi, Ali Al-Dahoud, Abdelfatah A Yahya “New outlier detection method based on fuzzy clustering”2011.

[5]. Deepak Soni, Naveen Jha, Deepak Sinwar,” Discovery of Outlier from Database using different Clustering Algorithms”, Indian J. Edu. Inf. Manage., Volume 1, pp 388-391, September 2012.

[6]. Han & Kamber & Pei,” Data Mining: Concepts and Techniques (3rded.) Chapter 12, ISBN-9780123814791

[7]. Ji Zhang,” Advancements of Outlier Detection: A Survey”, ICST Transactions on Scalable Information Systems, Volume 13, pp 1-26 January-March 2013

[8]. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis,” On Clustering Validation Techniques”, Journal of Intelligent Information Systems, pp 107–145, January 2001.

[9]. Maria Kontaki, Anastasios Gounaris, Apostolos N. Papadopoulos, Kostas Tsichlas, Yannis Manolopoulos,” Continuous Monitoring of Distance-Based Outliers over Data Streams”, Proceedings of the 27th IEEE International Conference on Data Engineering , Hannover, Germany, 2011.

[10]. Moh'd belal al-zoubi1, ali al-dahoud2, abdelfatah a. yahya” New Outlier Detection Method Based on Fuzzy Clustering”

[11]. Mr Ilango, Dr V Mohan,” A Survey of Grid Based Clustering Algorithms”, International Journal of Engineering Science and Technology, Volume 2, pp 3441-3446, 2010.

[12]. Ms. S. D. Pachgade, Ms. S. S. Dhande,” Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, pp 12-16 June 2012.

[13]. Periklis Andritsos,” Data Clustering Techniques”, pp 1-34, March 11, 2002.

[14]. P. Murugavel, Dr. M. Punithavalli,” Improved Hybrid Clustering and Distance-based Technique for Outlier Removal”, International Journal on Computer Science and Engineering, Volume 3, pp 333-339, 1 January 2011.

[15]. Sivaram, Saveetha,”AN Effective Algorithm for Outlier Detection”, Global Journal of Advanced Engineering Technologies, Volume 2, pp 35-40, January 2013.

[16]. S.Vijayarani, S.Nithya,” Sensitive Outlier Protection in Privacy Preserving Data Mining”, International Journal of Computer Applications, Volume 33, pp 19-27, November 2011.

[17]. S.Vijayarani, S.Nithya,” An Efficient Clustering Algorithm for Outlier Detection”, International Journal of Computer Applications, Volume 32, pp 22-27, October 2011

[18]. Silvia Cateni, Valentina Colla ,Marco Vannucci Scuola Superiore Sant Anna, Pisa,” Outlier Detection Methods for Industrial Applications”, ISBN 78-953-7619-16-9, pp. 472, October 2008

[19]. Shalini S Singh, N C Chauhan,” K-means v/s K-medoids: A Comparative Study”, National Conference on Recent Trends in Engineering & Technology, May 2011.

[20]. Tan, Steinbach, Kumar,” Introduction to Data Mining (1sted.) chapter 10”, ISBN-0321321367.