Efficient Clustering Algorithm for Outlier Detection

Dhananjay Kumar Tiwari[1], Hari Mohan Singh[2], Shiv Pratap Pal[3]

[1] *Department Computer Science & I.T., Sam Higginbottom Institute of Agriculture Technology & Sciences, Naini, Allahabad, U.P, India*
dhananjay.tiwari@shiats.edu.in

[2] *Department Computer Science & I.T, Sam Higginbottom Institute of Agriculture Technology & Sciences, Naini, Allahabad, U.P, India*
hari.singh@shiats.edu.in

[3]*Kendriya Vidyalaya No. 1, Akhnoor (J & K), India*
shivppal@gmail.com

Abstract— **Data mining helps to extract important and valuable knowledge from large massive collection of data. Several techniques and algorithms are used for extracting the hidden patterns from the large data sets and finding the relationships between them. Clustering algorithms are used for grouping the data items based on their similarity. Clustering means the act of partitioning an unlabeled dataset into group of similar objects. The goal of clustering is to group sets of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate clusters. The algorithms used in this research work are PAM (Partitioning around Medoid), CLARA (Clustering Large Applications) AND CLARANS (Clustering Large Applications Based on Randomized Search) and also a new clustering algorithm ENHANCED CLARANS (ECLARANS) for detecting outliers. In order to find the best clustering algorithm for outlier detection several performance measures are used. The experimental result shows that the outlier detection accuracy is very good in the ECLARANS clustering algorithm as compared to the existing algorithms. It has a very high accuracy but still it takes time to be accurate. The aim of this research is to reduce the time complexity of the ECLARANS and making it more powerful and effective method.**

Keywords— **DataMining, Clustering, PAM, CLARA, CLARANS and ECLARANS, Outlier Detection.**

I. INTRODUCTION

Data mining is becoming an important tool to convert the data into information. It is commonly used in a wide series of profiling practices, such as marketing, fraud detection and scientific discovery. 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. The mining process will be ineffective if the samples are not good representation of the larger body of the data. The discovery of a particular pattern in a particular set of data does not necessarily mean that pattern is found elsewhere in the larger data from which that sample was drawn. 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 and the measurement and is calculated through distance function. Clustering is a useful technique for the discovery of data distribution and patterns in the underlying data. It is an unsupervised learning technique. Unsupervised learning is learning from observations and discovery. In this mode of learning, there is no training set or the prior knowledge of the classes. The system analyses the given set of data to observe similarities emerging out of the subsets of the data. The outcome is a set of class descriptions, one for each class, discovered in the environment.

The aim of clustering analysis is to find any interesting groupings of the data. It is possible to define cluster analysis as an optimization problem in which a given function consisting of within cluster similarity and between clusters dissimilarities needs to be optimized. 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. Outlier detection has direct applications in a wide variety of domains such as mining for anomalies to detect network intrusions, fraud detection in mobile phone industry and recently for detecting terrorism related activities. Figure-1 depicts the general block diagram of outlier detection model. Some of the outlier detection techniques are:

- ✓ Distance based outlier detection
- ✓ Distribution based outlier detection
- ✓ Density based outlier detection
- ✓ Depth based outlier detection

Distance based outlier detection –In a distance based outlier detection a measure is given on a space, a point p in a dataset is considered outlier with respect to parameters M and D. The measure given by user, if it is less than M points within the distance D from p.

Distribution based outlier detection-A statistical model is decided and objects are applied to that model thereafter it is decided whether the point or object is coming into the statistical model or not.

Density based outlier detection-That objects falling in the low density region are considered to be outliers.

Clustering based outlier detection-In this approach various algorithms are used and by the help of them small and large clusters are made and small clusters are declared as outliers. Each of these techniques has its own advantages and disadvantages. In all these methods generally, the technique to detect outliers consists of two steps. The first identifies an outlier around a data set using a set of inliers (normal data). In the second step, a data request is analyzed and identified as outlier when its attributes are different from the attributes of inliers. The methodology is tested on the problem of cleaning official statistics data. Indeed, for cluster analysis to work effectively, there are the following key issues:

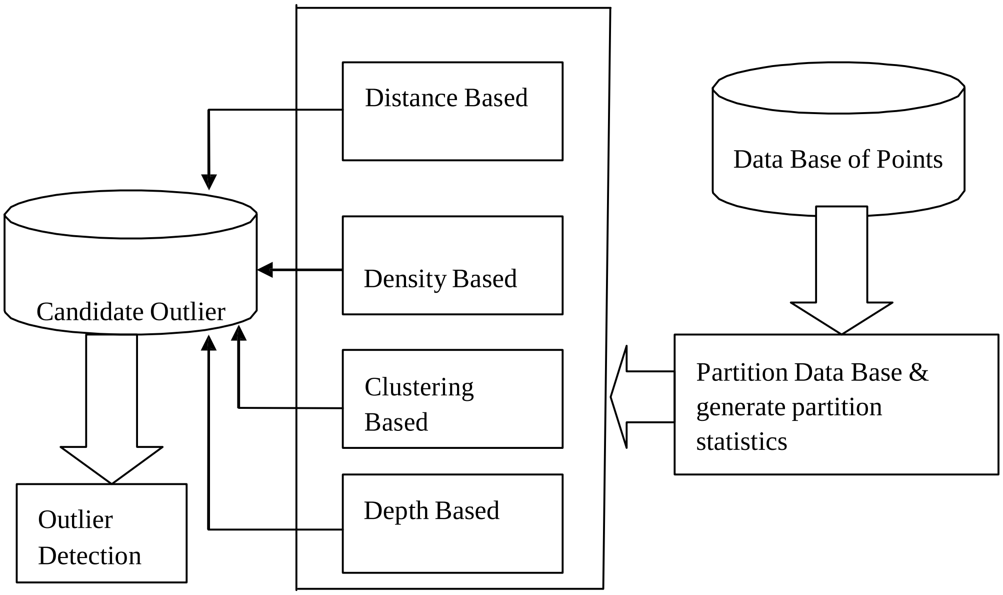

Fig. 1. Clustering based Outlier Detection

1. Whether there exists a natural notion of similarities among the “objects” to be clustered.
2. Whether clustering a large number of objects can be efficiently carried out. Traditional cluster analysis algorithms are not designed for large data sets, with say more than 1,000 objects

II. METHODOLOGY

The problem statement of this research work is to find out the outliers using ECLARANS algorithm (which is a partioning based clustering algorithm) with reduced time complexity. Along with that different clustering algorithms will be applied to compare and verify their performance.

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. The present study analyzes the use of PAM , CLARA, CLARANS and ECLARNS.

ENHANCED CLARANS (ECLARANS): This method is different from PAM, CLARA AND CLARANS. This method tends to improve the accuracy of outliers. ECLARANS is a partitioning algorithm which is an improvement of CLARANS to form clusters by selecting proper nodes instead of selecting the nodes through random search operations. The algorithm is similar to CLARANS but these selected nodes reduce the number of iterations of CLARANS ECLARANS Procedure. The Previous researches has successfully established ECLARANS as an effective algorithm for outlier detection but up till now it doesn’t have better time complexity. This research work tends to achieve better time complexity by using modified ECLARANS method.

Existing ECLARANS 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
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.
```

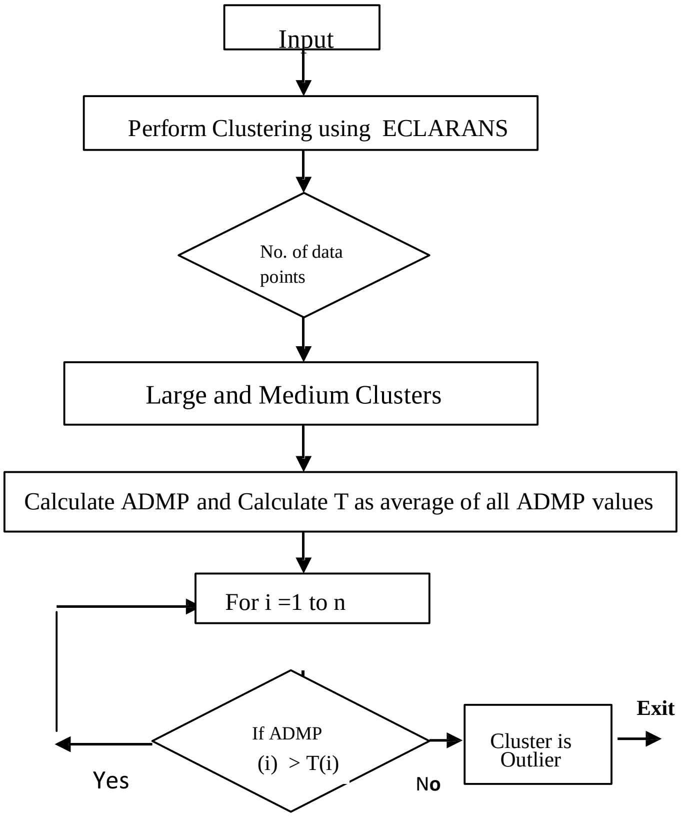
Input
Perform Clustering using ECLARANS
No. of data points
Large and Medium Clusters
Calculate ADMP and Calculate T as average of all ADMP values
For i =1 to n
If ADMP (i) > T(i)
Yes
No
Cluster is Outlier
Exit

Fig. 2. Flowchart- Procedure of Outlier detection

Modified ECLARANS 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 have 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 Step 4.

III. EVALUATION RESULTS

Proposed algorithm has been implemented using programming language Java and WHO dataset. The java has been used using Netbeans 7.3.1 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 executing the proposed algorithm the results obtained are tabulated in the following tables.

Outlier detection time (in second) for 8000 data 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

Outlier detection time (in second) for 4000 data 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

Outlier detection time (in second) for 2000 data objects

No. of execution	PAM	CLARANS	ECLARANS	Modified ECLARANS
1	10.255	10.195	10.125	10.065
2	10.464	10.384	10.334	10.254
3	10.004	9.924	9.874	9.794

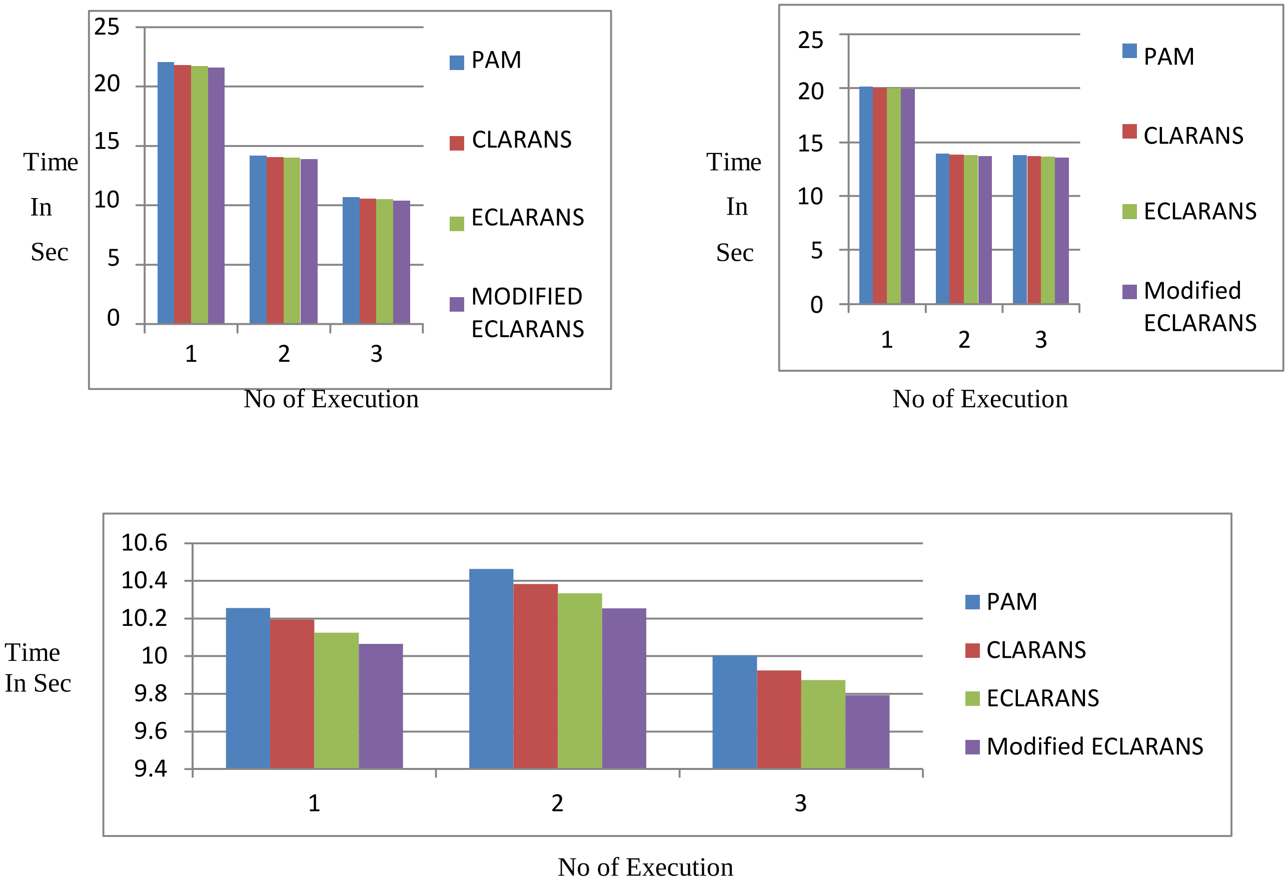

Fig. 3: Outlier detection time (in seconds) for 8000, 4000 and 2000 data objects

By the above charts it has been confirmed that the Modified ECLARANS is taking less time than the other algorithms. Even though time is varying with every execution due to number of threads executing each time on system still the difference between execution time of the entire algorithm is notable. Base paper takes all of the data points for calculating the cost and with that maximum cost it chooses the arbitrary data points as medoids where as in this new modified algorithm those data points which causes maximum cost are chosen instead of choosing the random data points.

IV. CONCLUSION

Modified ECLARANS has been found to be more accurate and time efficient as compared to their predeccassors namely PAM, CLARANS & ECLARANS. The comparison results are shown in the result analysis. Large number of Partition based outlier detection techniques are available, which can be used to solve all the problems. But all the algorithms are designed under certain assumptions and different algorithms are used under different conditions. Such as k-mean is used to handle spherical shaped cluster but we cannot use it 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 this 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 clusterin*g"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.*