# Enhancement of CURE Clustering Technique in Spatial Data Mining Using Oracle 11G

[1]Snehlata Bhadoria, [2]U. Datta
[1,2]Dept. of Computer Science and Engineering, MPCT, Gwalior, Madhya Pradesh, India

**Abstract**
CURE Clustering divides the data sample into groups by identifying few representative points from each group of the data sample. This paper presents enhanced CURE as a clustering technique for data mining, in this approach we have a specially designed pattern as representative to form enhancement in CURE clustering to make it more usable efficiently on big data. Oracle 11G is used as backend with its silent feature of storing big data. The supervised trained model used to analyze this pattern to enhancing the CURE clustering which execute their function by specified parameter or value. This algorithm makes clustering easier and applicable on huge data by reducing time complexity.

**Keywords**
CURE, Soft Pattern Analysis

## I. Introduction

In this paper, the proposed technique is highly inspired by the CURE of hierarchical clustering.

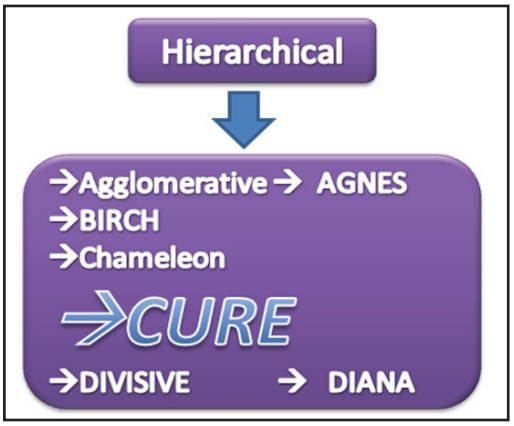



Fig. 1: CURE & Other types of Hierarchical Clustering

CURE employs features of both the centroid based algorithms and all algorithms point [8]. CURE obtains data sample from the given database. The algorithm CURE divides data sample into number of groups and identifies representative points from each group of the data sample. In the first phase of algorithm a set of widely spaced points from the given datasets is considered. In the next phase selected dispersed points are moved to centre of the cluster by a specified value of a factor α. As result of this process few randomly shaped clusters are obtained from datasets. In the process it identifies the eliminated outliers. In the next phase of the algorithm, the representative points of the clusters are checked for proximity with threshold value and the clusters that are next to grouped together to form the next set of clusters. In this hierarchical algorithm the value of factor α may vary between 0 and 1. The utilization of shrinking factor α overcomes the limitations of centroid based and all-points approaches by CURE. As the representative points move through clustering space, the ill effects of outliers are reduced by a greater extent. Thus the shrinking factor α enhance the feasibility of CURE. The worst case time complexity of CURE is determined to be O ($n^2 \log n$).

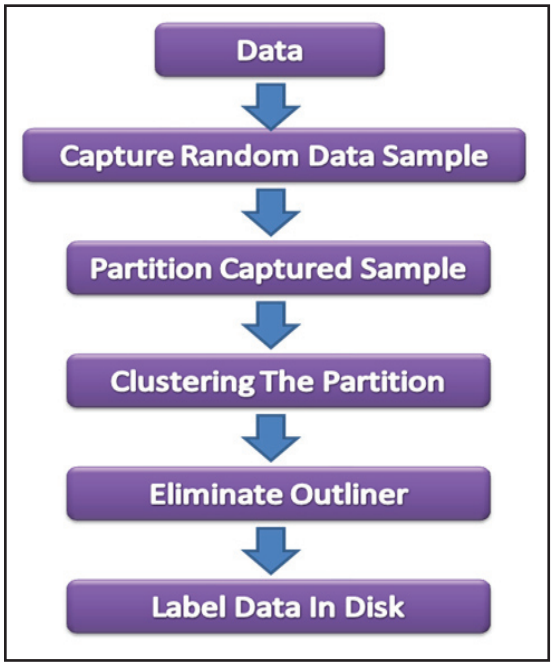



Fig. 2: Steps of CURE

Fig. 2 shows the overview of CURE implementation. A random sample of data objects is drawn by datasets. Partial clusters are obtained by partitioning the sample dataset and outlier identified and removed by this stage. Final refined clusters are formed from the partial cluster set.

## II. Issues and Analytic Report on Previous Algorithm

### A. Major Issues

The mostly observed issues in data mining limited for handling large databases. Some of them are listed below.

#### 1. Issues in Mining Technique and User-interaction

- These deem the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad-hoc mining, and knowledge visualization
- Performance issues: In this efficiency, scalability, and parallelization of data mining algorithms are considered.
- Issues relating to diversity of database types: Insertion, deletion updating anomalies comes in this category

### B. Just Previous Task

To improve the efficiency and performance of clustering with reducing issues probability a generalized parallelization algorithm in CURE clustering used.

**Design of Parallel algorithm** Basic Techniques of Parallel Processing various programming models of parallel processing are characterized by parallelism used and type of inter-processor communication used.

Data divides into partitions, so also the task into subtasks. Accordingly, based on whether a data is divided into partitions

three kinds of parallelism can be identified viz. Data Parallelism, Task Parallelism and Hybrid Parallelism. Based on different processors communication, there are three kinds of parallel processing systems, Shared Memory Systems, Message Passing Systems and Remote Memory Operations System.

### 1. Data Parallelism
The number of partitions made from original data and the same program runs on each of the partitions. The results program obtained on each data partitions is later combined to get the final result.

### 2. Task Parallelism
Main task is divided into set of smaller subtasks which are identified and assigned to individual processors. The independent processors are used to get the final solution. This speeds up the execution by multifold as the tasks are performed concurrently.

### 3. Hybrid Parallelism
Both data and tasks are divided. The subtasks that can be perform on independent data partitions, whose results can be merged together, are identified along with the independent subtasks. These subtasks applied to independent processors in proper sequence and using the results of these parallel processing, final result is obtained.

### 4. Shared Memory Systems
All the processors share a common global memory. Locking of shared memory for two or more processors those try to use the shared memory for writing.

### 5. Message Passing
Each processors work on its own memory. The processors communicate by sending and receiving the messages explicitly using send and receive commands.

### 6. Remote Memory Operations
Processors can explicitly access memory of other processors. Individual commands accessing local and remote memory. Parallel computing allows using multiple CPUs where a problem is broken into various parts which could be solved concurrently. Each part is further broken to a series of instructions and instructions from every part execute concurrently on different CPUs. Before making new algorithm for existing CURE clustering technique, we have a look to common data mining algorithm,
such as :

```
{//Outer Sequential Loop}
While()
{
        //Reduction Loop
        For (element e)
        {
                (i, val) = process( e ) ;
                Reduc(i) = Reduc(i) op val ;
        }

}
```

## C. Parallel Architectures
A parallel computer or multi-processor system is a computer utilizing more than one processor. A common way to classify parallel computers is to distinguish them by the way how processors can access the system's main memory because this influences heavily the usage and programming of the system.

**Parallel Performance Analysis** Performance analysis is an iterative subtask during program development. The goal is to identify regions that do not perform well. Performance analysis is structured into three phases:

### 1. Measurement
Performance analysis is done based on information on runtime events gathered during program execution. The basic events are, for example, cache misses, termination of a floating point operation, start and stop of a subroutine or message passing operation.

### 2. Analysis
During analysis the collected runtime data are inspected to detect performance problems. Performance problems are based on performance properties, such as the existence of message passing in a program region, the programmer applies a threshold. Only performance properties whose severity exceeds this threshold are considered to be performance problems.

### 3. Ranking
During program analysis the severest performance problems need to be identified i.e. problems need to be ranked according to the severity. Current techniques for performance data collection are profiling and tracing. Profiling collects summary data only. This can be done via sampling. Tracing is a technique that collects information for each event.

## D. Step- By – Step Analysis
The designed parallel algorithm used as CURE for data parallelism and shared memory systems in CRCW PRAM. The measured elapsed time, speedup and scale up of CURE. Speedup gives efficiency of parallel algorithm during change in number of processors. Another interesting measure is scale up. Scale up captures how a parallel algorithm handles larger datasets when more processors are available. It is observed that the elapsed times of CURE on different numbers of processors. Total execution time substantially decreases as the number of used processors increases. In particular, for the largest datasets the time decreases in a more significant way. It could be observed as the dataset size increases the time gain increases as well for speedup results obtained for different datasets, and can observe that the CURE algorithm scales well up to 12 processors for the largest datasets, whereas for small datasets the speedup increases until the optimal number of processors are used for given problem, e.g., 3 processors for 5000 tuples or 7 processors for 20000 tuples. When more processes are used the algorithm does not scale because processors are not effectively used and communication costs increases.

Table 1: Process Manipulation With Used Processor

| S. No. | Processor Used | Number of manipulated tuples |
|---|---|---|
| 1. | 03 | 5000 |
| 2. | 07 | 20000 |
| 3. | 12 | 35000 |

Hierarchical clustering algorithms repeatedly split or aggregate data by hierarchical structure, in order to form a hierarchical sequence of solutions. The complexity is $O(n2)$, and applicable to small scale dataset. For example, CURE [16] uses a novel hierarchical strategy through selection of fixed number of representative points and multiplying a shrinking factor to approach the centre of the cluster.

## III. Proposed Enhanced Cure Algorithm

Designed patterned human unique ID, This patterned is an example on which our proposed approach of enhanced CURE on huge data of Oracle 11G applied. Here the component of generated pattern worked as representative to perform clustering by trained model of CURE to increase the time efficiency of Clustering.

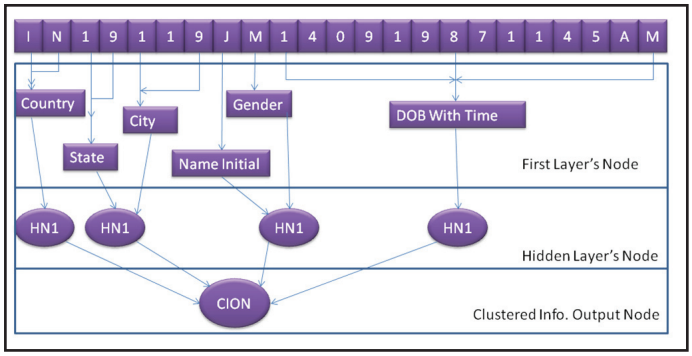



Fig. 3: Demonstration of Drown Pattern to Show Trained Model Working of Enhanced CURE.

In the proposed modal the CURE algorithm enhanced by the designed trained model to analyze the drown pattern to perform clustering faster than available techniques.
As the above specified pattern which could also generate with more attributes to uniquely identify the county people. The proposed model could work on any specifically design pattern to perform clustering by our enhanced CURE approach. Steps of ours task's manipulation and flow of instruction is shown below.

**Step 1:** The Designed pattern apply as input to first layer of our trained model, the model will change their priority of attribute selection dynamically.
**Step 2:** The made supervised trained model is apply to enhancing the CURE by its fastest processing among layers and nodes.
**Step 3:** The hidden layer manipulate the assigned way of clustering.
**Step 4:** The out is as clustered information of entities in BIG DATA.

### A. Proposed Enhanced Cluster Algorithm

The assign pattern will work as input to trained system of etrievi CURE.

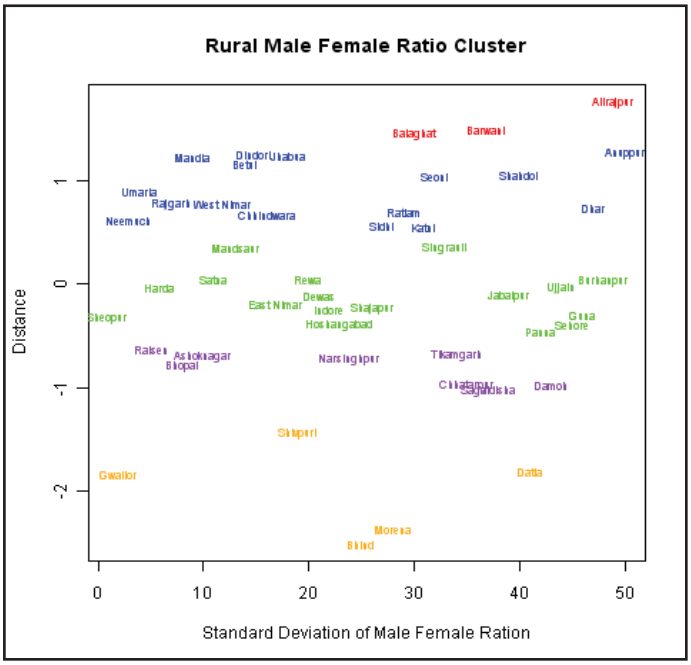
```
Step 1:  Get Input String Pattern
         GET_INPUT()
         {
         Select patterned string by well known oracle query.
         Select.
         } // Made Get Input Function

Step 2:  Get Partitioned Information by first procession phase.
         GET_Partitioned_Info();
         {
         For(i=0;i>=0&& I <=patern id lenght)
         Get = text[i]
         }

Step 3: Analyze();
         {
         Get
         country = f1; state =f2; city = f3; name initial =f4;
         gender =f5; DOBWT=f6
         enhanced_trained_cure(f1,f2,f3,f4,f5,f6);
         //Get info by trained model of enhanced CURE.
         }
         // Apply Enhanced CURE to Analyze partitioned info
         by trained system to get valuable component
         as Clustered Info.

Step 4: get_info();
         {
         Getinfo(f1||f2||f3||f4||f5||f6)
         //Show final clustered date by available aspects. ;
         }
         // get clustered information by got

Step 5: End
```

## IV. Result and Comparison

The observed result is in the form of cluster by selected attribute and another output is to gain like the ratios of male female and population growth with respect to time. As shown below in male, female ration of graphical analysis of created sub cluster of MP state. Table 2 representing the competitive analysis with made cluster approaches. This could perform well as on one processor same on other number of processor with equitation ratio of available data or machine configuration.

Table 2: Comparative Analysis

| Algorithm | Type of Clustering | Dimensionality of Data | Parameters Used | Shape of the clusters | Worst Case Time |
|---|---|---|---|---|---|
| OPTICS | Global Clustering and Local Clustering | Multidimensional data, large datasets | Multiple number of distance parameter ε | Arbitrary Shapes | O(n) |
| ROCK | Local | Small Sized | Similarity Threshold | Graph | O(n2) |
| CHAMELEON | Global Clustering and Local Clustering | Multidimensional data, large datasets | relative inter-connectivity RI (Ci, Cj), relative closeness RC (Ci, Cj) | Arbitrary Shapes | O(nm+nlogn+m2logm) |
| CURE | Global Clustering and Local Clustering | Multidimensional data, large datasets | shrinking factor α | Nonspherical | WC $O(n^2 \log n)$<br>AC $O(n^2)$<br>BC O(n) |
| Proposed Enhanced CURE | Global Clustering and Local Clustering | Multidimensional data, large datasets | shrinking factor α (Dynamic) | Nonspherical | WC O(nlog n)<br>AC O(n)<br>BC O(n-2) |




Graph 1: Standard Deviation of Male Female Ration

## V. Conclusion

Here presented enhanced CURE technique reduces effort and time with respect to just previous and series of efforts putted on it. To make it better the algorithm completed their task within four steps as shown in proposed architecture of enhanced CURE. This could be performing on any huge backend like taken oracle 11G or rest other. In the next effort we will use dynamic learning with currently proposed technique which would be better then in time saving and would be applicable on any kind of data like images and videos.

## References

[1] Smiti, Abir, Zied Eloudi,"Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory", In the IEEE 6th International Conference on Human System Interaction 2013, pp. 380-385, 2013.

[2] Neha Soni, Amit Ganatra,"Categorization of Several Clustering Algorithms from Different Perspective: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 8, pp. 63-68, Aug. 2012.

[3] Pragati Shrivastava, Hitesh Gupta,"A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research, pp. 2249-7277, September 2012.

[4] Chaudhari Chaitali G.,"Optimizing Clustering Technique based on Partitioning DBSCAN and Ant Clustering Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) Vol. 2, Issue 2, pp. 212 – 215, December 2012.

[5] Chakraborty S., Prof. Nagwani N. K.,"Analysis and Study of Incremental DBSCAN Clustering Algorithm", International Journal of Enterprise Computing And Business Systems, Vol. 1, July 2011.

[6] Chandra. E, Anuradha. V. P,"A Survey on Clustering Algorithms for Data in Spatial Database Management System", International Journal of Computer Applications, Vol. 24, June 2011.

[7] Parimala M., Lopez D., Senthilkumar N. C.,"A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology, Vol. 31, June 2011.

[8] Santhisree K., Dr. Damodaram A.,"SSMDBSCAN and SSM-OPTICS : Incorporating new similarity measure for Density based clustering of Web usage data", In International Journal on Computer Sciences and Engineering, August 2011.

[9] Anant Ram, Sunita Jalal, Anand S. Jalal, Manoj Kumar, "A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases", International Journal of Computer Application Vol. 3, No. 6, June 2010.

[10] Dr. E. Chandra, V. P. Anuradha," A Survey on Clustering Algorithms for Data in Spatial Database Management Systems", International Journal of Computer Application, Vol. 24, pp. 19-26.

[11] A. K. Jain,"Data Clustering: 50 Years Beyond K-Means, in Pattern Recognition Letters, Vol. 31 (8), pp. 651-666, 2010.

[12] Peter J. H., Antonysamy A.,"An optimized Density based Clustering Algorithm", International Journal of Computer Applications, Vol. 6, September 2010.

[13] Ram A., Jalal S., Jalal A. S., Kumar M.,"A Density based Algorithm for Discovering Density varied clusters in Large Spatial Databases", International Journal of Computer Applications, Vol. 3, June 2010.

[14] Jitendra Singh Sengar,"Soft Computing Approach to Evaluate Trust Value of a Node in VNET with HMM through Supervised Learning", International Journal of Communication Technology for Social Networking Services Vol. 1, No. 1, 2013, pp. 11-14.

Snehlata Bhadoria received her B.E. degree in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior M.P., in 2001. She worked as a lecturer, with Information Technology department in Rustam Ji Institute of Technology BSF Tekanpur and with Computer Science department in Shri Ram College of Engineering & Management. Banmore, Gwalior. She is pursuing M.E. degree from MPCT Collage in Computer science and Engineering, Gwalior. Her research interests include Enhancement of CURE clustering technique in Spatial Data mining using Oracle 11G.