ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

Grid Density Clustering Algorithm

Amandeep Kaur Mann[1], Navneet Kaur[2],

Scholar, M.Tech (CSE), RIMT, Mandi Gobindgarh, Punjab, India [1]

Assistant Professor (CSE), RIMT, Mandi Gobindgarh, Punjab, India[2]

Abstract: Data mining is the method of finding the useful information in huge data repositories. Clustering is the significant task of the data mining. It is an unsupervised learning task. Similar data items are grouped together to form clusters. These days the clustering plays a major role in every day-to-day application. In this paper, the field of KDD i.e. Knowledge Discovery in Databases, Data mining, clustering analysis and the prevailing the Grid Density Clustering Algorithm are described.

Keywords: Data mining, KDD, Clustering, Cluster Analysis, Grid Density Clustering Algorithm**.**

I. INTRODUCTION

KDD stands on the Knowledge Discovery in Databases is the process of finding associations and patterns in raw data automatically from large databases and gives the output results.

In particular, the KDD process consists of the following steps:

1. Selection and Integration: The data that is to be mined may exist in dissimilar and mixed data form. Therefore, It is necessary to first selecting the appropriate data from a variety of databases for analysis and integrate that data values into a logical data store.

2. Pre-processing: Raw data may have invalid or missing values. Invalid values are corrected while misplaced values are supplied or predicted.

3. Transformation: The data are transformed into representations that are suitable for mining tasks.

4. Data Mining: Data Mining is the core part of the KDD process referring to the purpose of intelligent techniques to take out the secreted knowledge from the transformed data.

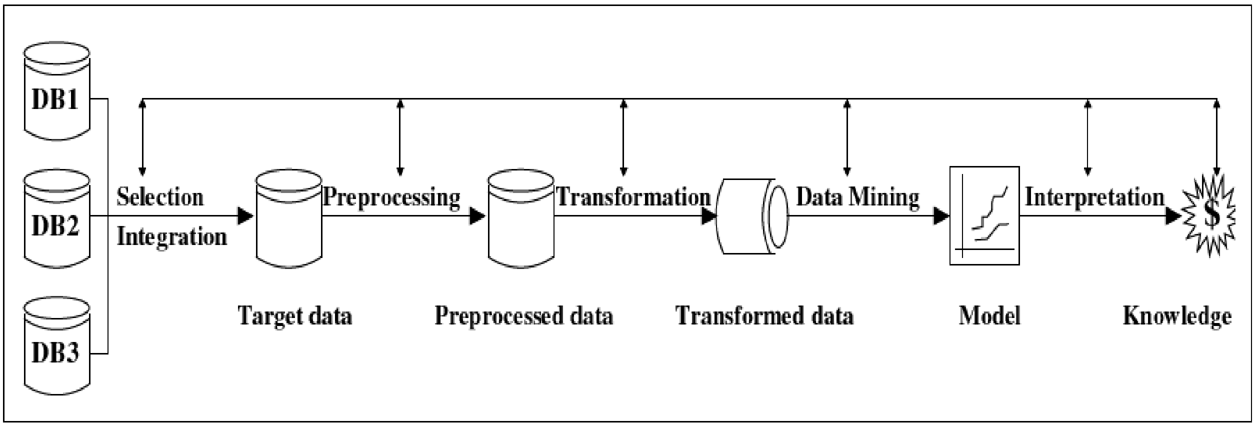
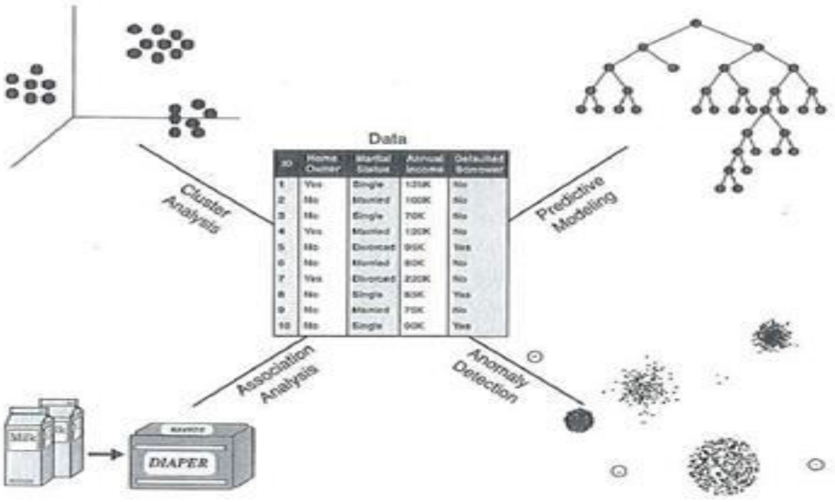


Figure1: Steps of the KDD process.

5. Interpretation and Evaluation: A data mining system has the ability to produce a great number of patterns but only a small fraction of them may be of attention. Therefore the suitable methods are required to estimate the valuable results.

Data Mining

Data mining is the method of without human intervention for finding the useful information in huge data repositories. Data mining involves six common tasks: Anomaly Detection, Association rules learning, Classification, Clustering and Regression. In this paper we only discussed the clustering. Clustering is the most significant unsupervised-learning problem. The major function of clustering is to finding a structure of similar data items. Totally, the clustering involves partitioning a given dataset into a number of groups of data whose members are similar in some way. The usability of cluster analysis has been used broadly in data recovery, pattern recognition, text and web mining, software reverse engineering and image segmentation.



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 2: The four core of Data mining task

Specific uses of data mining include:

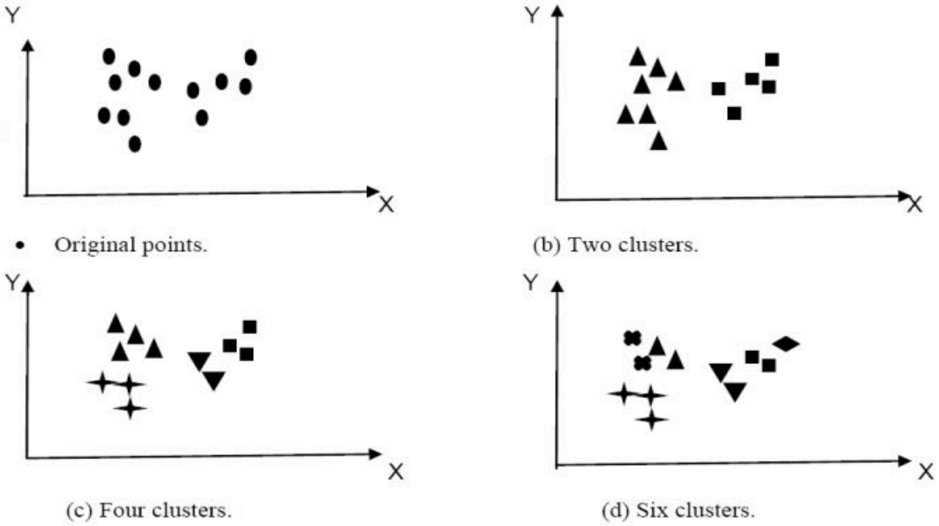
- Market segmentation - Identify the universal characteristics of customers who buy the similar products from your company.
- Customer churn - calculate which customers are likely to go away from your company and buy the products from your competitor.
- Fraud detection - Identify which transactions are most likely to be fraudulent.
- Direct marketing - Identify which prediction should be included in a mailing list to gain the maximum response rate.
- Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
- Market basket analysis - Understand and identify the products or services that are usually purchased together; e.g., soap and oil.
- Trend analysis - Reveal the difference between typical customers this month and last

.

Dividing the data objects in significant groups of data objects or classes known as cluster are based on familiar attribute and play an important role in how people analyse and explain the world. For an example, children can speedily label the object in a photograph, such as buildings, trees, people and so on. In the field of understanding data, clusters are potential classes, and cluster analysis is a studying method to discover classes.

Cluster Analysis

Cluster Analysis method as fields grow extreme rapidly with the objective of together data objects, based on information found in data and describing the associations inside the data. Cluster analysis is a descriptive data analysis task that aims to find the intrinsic structure of a collection of data points (or objects) by partitioning them into identical clusters based on the values of their attributes. A similarity metric is defined between the data objects, and then similar objects are grouped together to form clusters. Clustering is unsupervised learning since it does not require assumptions about category labels that tag objects with prior identifiers. Since there is no universal definition of clustering, there is no universal measure with which to evaluate clustering algorithms.



- Original points.

(b) Two clusters.

(c) Four clusters.

(d) Six clusters.

Figure 3: Different types of clusters

II. GRID DENSITY BASED ALGORITHMS

Grid Density based clustering is concerned with the value space that surrounds the data points not with the data objects. This algorithm uses the grid data structure and use dense grids to form clusters. It first quantized the original data space into finite number of cells which form the grid structure and then perform all the operations on the quantized space. Grid based clustering maps the infinite amount of data records in data streams to finite numbers of grids. Its main uniqueness is the fastest processing time, since like data points will fall into similar cell and will be treated as a single point. It makes the algorithm self-governing of the number of data points in the original data set. The grid-based clustering algorithms are STING, Wave Cluster, and GDCLU. Grid density takes the advantage of the density and the grid algorithms. Grid density is suitable for handling noise. It can find the arbitrary shaped clusters used for high dimensional data. The grid density algorithm does not require the distance computation. K-mean knows the number of clusters in advance but the grid density does not. Grid density algorithm is better than the k-mean algorithm in clustering. The advantage of grid density method is lower processing time. Therefore, we implement the grid density clustering algorithm for analyse and increase the speed, and accuracy of the dataset. We implement this in the MATLAB environment. The name MATLAB stands for matrix laboratory. MATLAB is a numerical computing environment and fourth-generation programming language. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. The steps of the grid based algorithm are:

1. Creating the grid structure, in other words divide the data space into a finite number of cells.
2. Calculating the cell density.
3. Sorting of the cells according to their densities.
4. Identifying cluster.

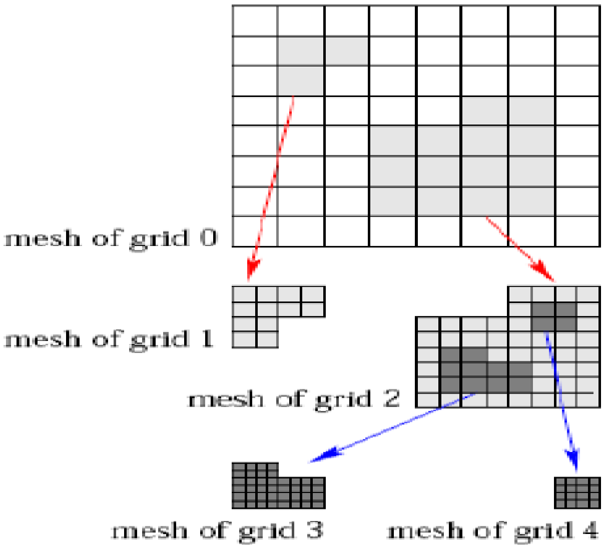


Figure4: Grid based Clustering

GDCLU

GDCLU generates major clusters by merging dense grids. Grid density is defined as number of points mapped to one grid. A grid is called sporadic when its density is less than the input argument. Two dense grids should be special neighbours in order to be merged; by merging these grids all their special neighbours also will be joined to the similar cluster. Based on the input parameter density, the algorithm is processed. The different types of the dataset are taken and their performance is analysed

III. RESULTS

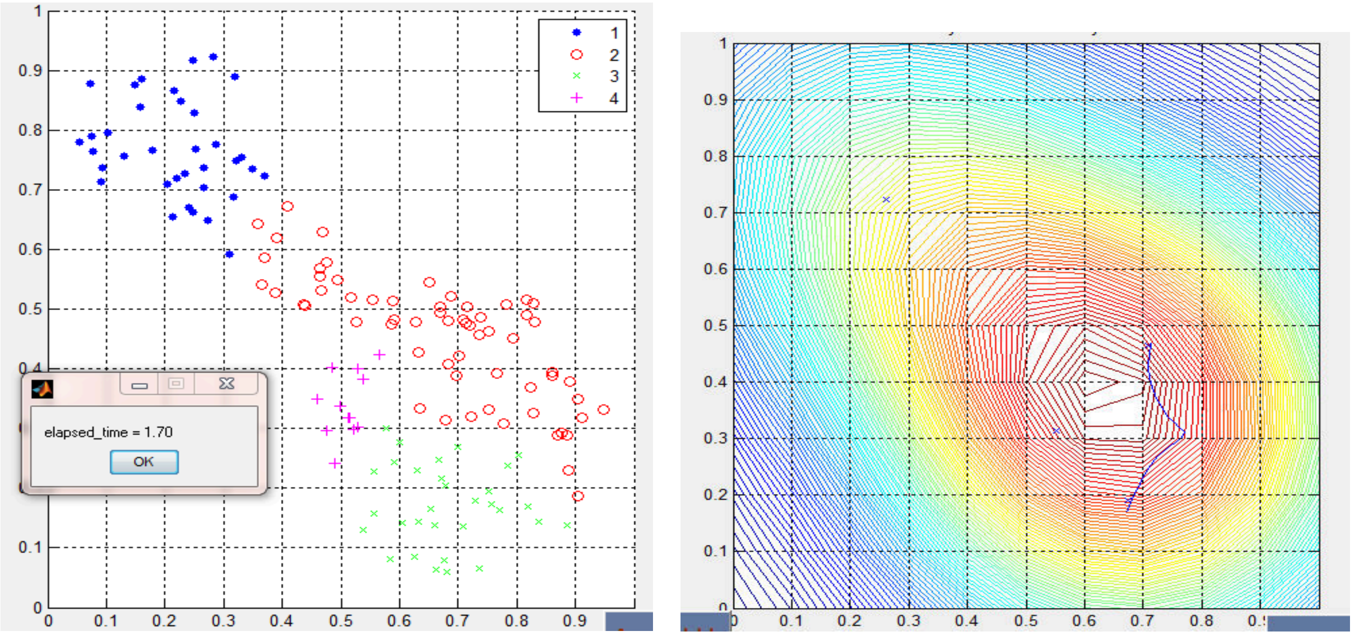
The results obtained from grid density clustering algorithm on different types of dataset based on number of numeric data values are shown in figure 5, 6, 7, 8. As the data set values are increases the computation time also increases. When the number of clusters is more than the execution time is also more.

The analytical results of the entire algorithms under the following parameters are implemented for given clustering algorithms and results are shown:

Elapsed Time: It is defined as the completion time of the algorithm. Lesser the elapsed time, less time to execute the algorithm more efficient is the algorithm. Elapsed time is calculated by measuring the finishing time of the exit task by the algorithm.

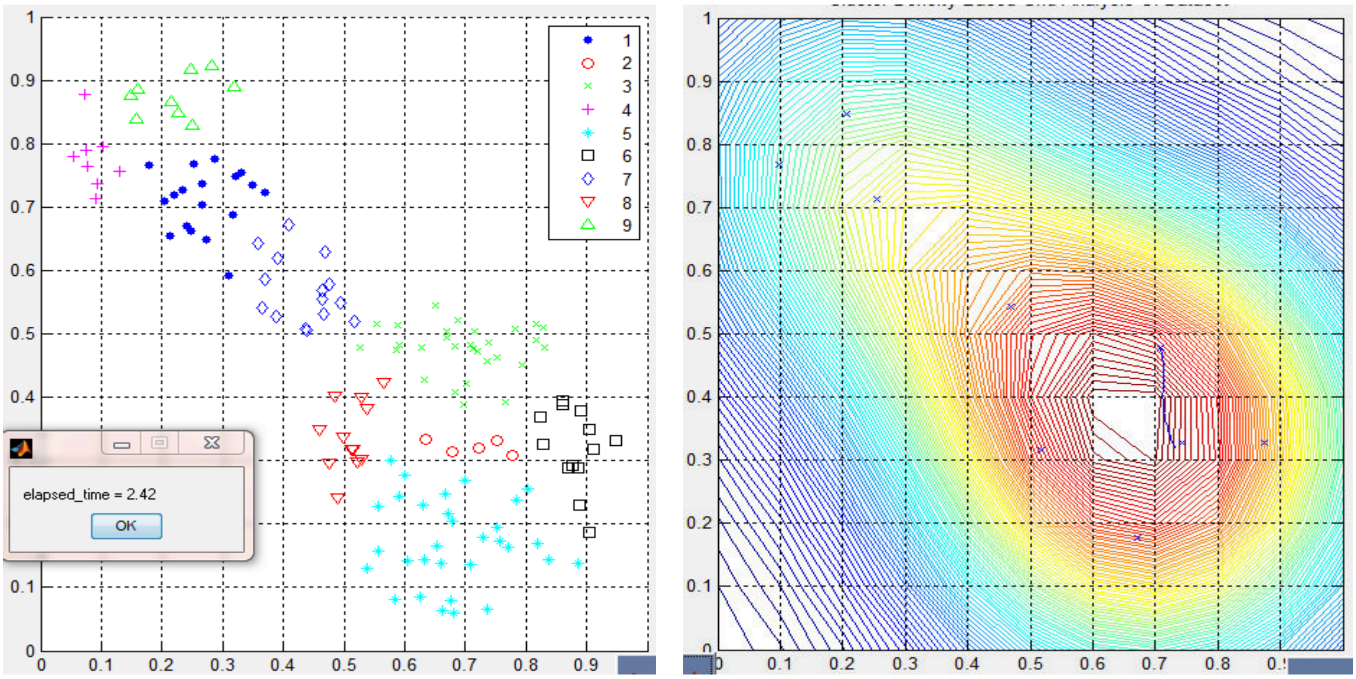
Clusters: Clusters defined as the similar data items in one group. More clusters means the more clearly representation of data values and easily findable. When more clusters are formed then elapsed time increases.

Density: Grid density is defined as number of points mapped to one grid. A grid is called sporadic when its density is less than the input argument



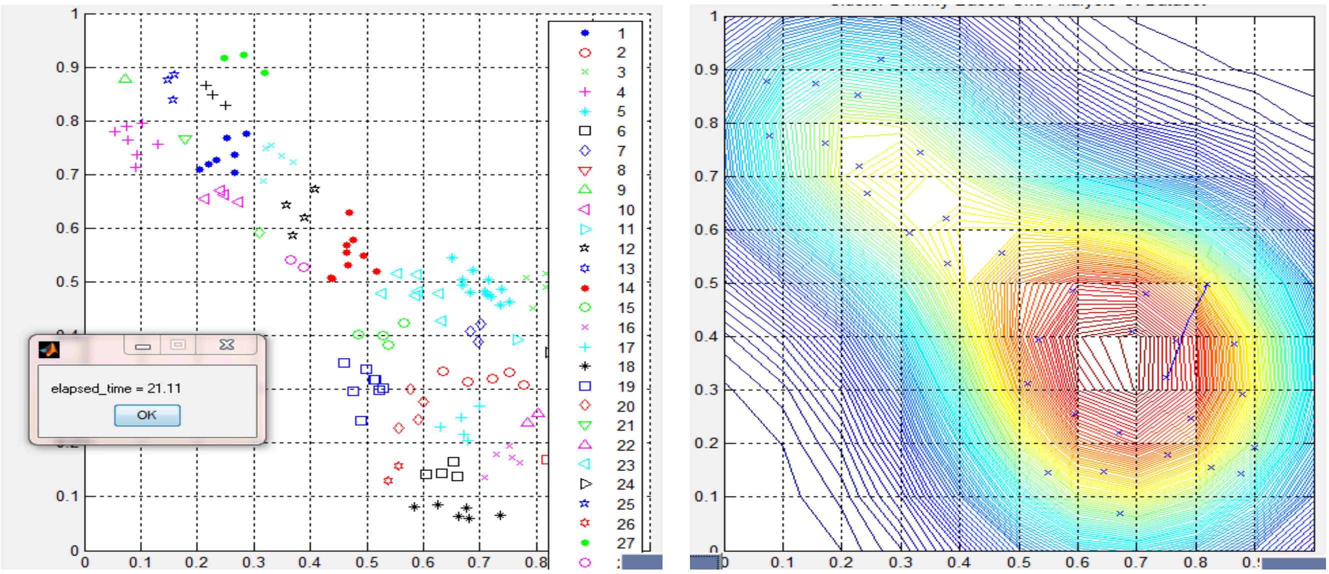
elapsed_time = 1.70

Figure5: Computation time and analysis of grid for 4clusters



elapsed_time = 2.42

Figure6: Computation time analysis of grid for 9clusters



elapsed_time = 21.11

Figure7: Computation time and analysis of grid for 28clusters

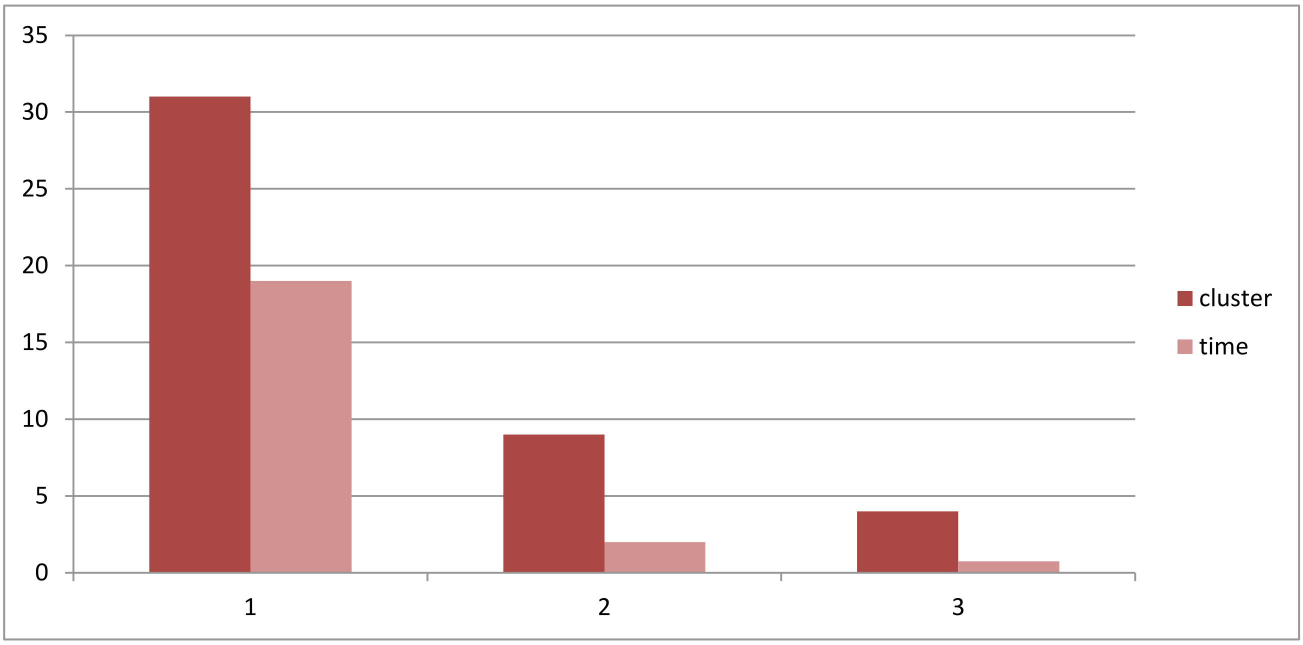


Figure8: This figure shows that as the accuracy of the data values are increased, and then the execution time is also increased.

IV. CONCLUSION AND FUTURE SCOPE

The clustering is the part of data mining. Clustering is used everywhere in our daily life. Traditional way of clustering is the partitional algorithm that is the k-mean. Although this algorithm is a number of limitations but this is being widely used in every part of the clustering and that clustering is the soft clustering. Grid density takes the advantage of the density and the grid algorithms. Grid density is suitable for handling noise. It can find the arbitrary shaped clusters. Grid density algorithm is better than the k-mean algorithm in clustering. The advantage of grid density method is lower processing time. The input parameters of any algorithms to get the best results are very carefully taken. This algorithm is implemented on the numeric dataset. The execution time depends upon the how large the size of the dataset. The future scope of this work is that we can also implement this hybrid approach clustering algorithm in the alphanumeric data and see the impact of the results and by varying the input parameters. As these results are concluded by using dense grids, the new approach can be developed by using the C-mean with Grid Clustering algorithm.

Acknowledgement

I would like to thanks the Department of Computer Science & Engineering of RIMT Institutes near Floating Restaurant, Sirhind Side, Mandi Gobindgarh-147301, and Punjab, India.

References

1. Wei-keng Liao, et.all. "**A Grid-based Clustering Algorithm using Adaptive Mesh Refinement",** Appears in the 7th Workshop on Mining Scientific and Engineering Datasets, pp.1-9, 2004.
2. Oded Maimon, Lior Rokach, "**DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK",** Springer Science+Business Media.Inc, pp.321-352, 2005.
3. GuiBin Hou, RuiXia Yao, et.all., **"Irregular Grid-based Clustering Over High-Dimensional Data Streams"** , IEEE 1st International Conference on Pervasive Computing, Signal Processing and Applications, pp.783-786, 2010.
4. MR ILANGO, Dr V MOHAN, "**A Survey of Grid Based Clustering Algorithms",** International Journal of Engineering Science and Technology, pp 3441-3446. 2010.
5. Zheng Hua, Wang Zhenxing, et.all, **"Clustering Algorithm Based on Characteristics of Density Distribution",** IEEE 2nd International Conference On Advanced Computer Control, vol.2**,** pp.431-435, 2010.
6. Amineh Amini, Teh Ying Wah, et.all**, "A Study of Density-Grid based Clustering Algorithms on Data Streams",**IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery, vol.3, pp.1652-1656, 2011.
7. Guohua Lei, Xiang Yu, et.all, **"An Incremental Clustering Algorithm Based on Grid",**IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp.1099-1103, 2011.
8. Amineh Amini1, Teh Ying Wah**, "DENGRIS-Stream: A Density-Grid based Clustering Algorithm for Evolving Data Streams over Sliding Window",** International Conference on Data Mining and Computer Engineering, pp-206-211, 2012.
9. Cheng-Fa Tsai, Tang-Wei Huang, **"QIDBSCAN: A Quick Density-Based Clustering Technique"**, IEEE International Symposium on Computer, Consumer and Control, pp.638-641, 2012