ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 6, June 2013

Grid Density Based Clustering Algorithm

Amandeep Kaur Mann, Navneet Kaur

***Abstract*— Clustering is the one of the most important task of the data mining. Clustering is the unsupervised method to find the relations between points of dataset into several groups. It can be done by using various types of techniques like partitioned, hierarchical, density, and grid. Each of these has its own advantages and disadvantages. Grid density takes the advantage of the density and the grid algorithms. In this paper, the comparison of the k-mean algorithm can be done with grid density algorithm. Grid density algorithm is better than the k-mean algorithm in clustering.**

***Index Terms*—Clustering, Data types, K-mean, Grid density.**

I. INTRODUCTION

Clustering is the one of the most important task of the data mining. Clustering is the unsupervised method to find the relations between points of dataset into several groups. Unsupervised method means the user known the input data but not known the output data but the way the user has to use for obtaining results is known. Cluster analysis can be done by Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. It is the unsupervised process that is why no predefined class is present. It has the major two applications: It is used as a stand-alone tool to get insight into data distribution; it is a preprocessing step for other algorithms. It has a rich applications and multidisciplinary efforts such as Pattern Recognition, Spatial Data Analysis in which it Create thematic maps in GIS by clustering feature spaces and detect spatial clusters or for other spatial mining tasks, Image Processing, Economic Science (especially market research), Document classification and Cluster Weblog data to discover groups of similar access patterns.[3]. Various examples of clustering are marketing, landuse, insurance, city planning, earthquake studies, transportation system etc. Generally the user has the large dataset say DataSet(X) when it is pass through the clustering algorithm it partitioned the DataSet(X) into P number of clusters.

As in the figure1, the dataset is passed through the clustering algorithm, then the dataset can be taken as output in the form of clusters, clusters are of arbitrary shaped clusters.

***Amandeep Kaur Mann**, M.Tech Computer Science & Technology, Pinjab Technical University/ RIMT/ RIMT Institutes of Technology, Patiala, India.*
***Navneet Kaur**, Assistant Professor Computer Science & Technology, Pinjab Technical University/ RIMT/ RIMT Institutes of Technology, Patiala, India.*

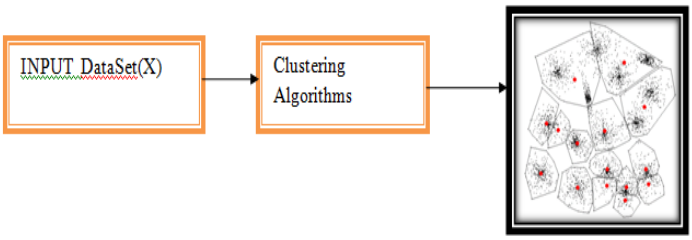

Figure1: Clustering Process

The quality of a clustering result depends on equally the similarity calculation used by the method and its implementation. The quality of a clustering technique is also calculated by its ability to discover some or all of the unknown patterns. Dissimilarity/Similarity metric is also used for this purpose. Similarity is spoken in conditions of a distance function, usually metric: $d(i, j)$. There is a separate "quality" function that calculates the "goodness" of a cluster. The definitions of distance functions are generally very different for interval-scaled, Boolean, categorical, ordinal ratio, and vector variables. Weights should be associated with dissimilar variables based on applications and data semantics. It is tough to define "similar enough" or "good enough", the answer is typically highly subjective. Some of the requirements of clustering in data mining are scalability, capability to deal with dissimilar types of attributes, capability to hold dynamic data, detection of clusters with arbitrary shape, negligible requirements for domain knowledge to determine input parameters, capable to deal with noise and outliers, tactless to order of input records, High dimensionality, merging of user-specified constraints, interpretability and usability. A good clustering algorithm produce high superiority clusters with low intra-cluster variance and high inter-cluster variance. In other words high similarity between intra-class and low similarity between inter-class clusters.

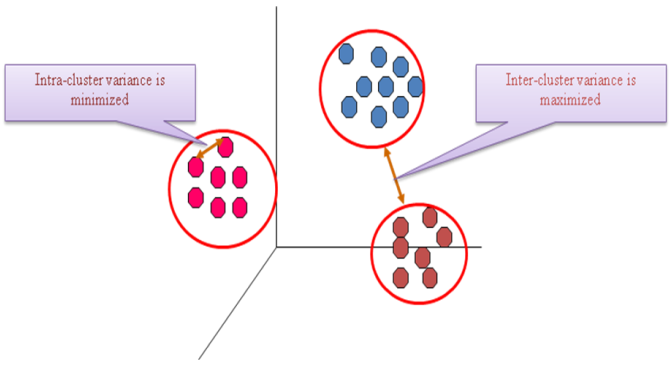

Figure2: Intra-class and Inter-class variance

Different types of data that is used for cluster analysis are [3]:

1. Interval-valued variables: from the standardize data calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

Where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$$

2. Binary variables: It is the contiguous table for binary values.
3. Nominal variables: A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green.
4. Ordinal variables: An ordinal variable can be discrete or continuous, Order is important, e.g., rank.
5. Ratio-scaled variable: It is a positive measurement on a nonlinear scale, approximately at exponential scale, such as Ae^{Bt} or Ae^{-Bt} .
6. Variables of mixed types: A database may contain all the six types of variables, symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio.

There are different algorithms for clustering. Some of the clustering algorithms require that the number of clusters should be known earlier to the start of clustering method others find out the clusters themselves. K-mean known the number of clusters earlier usually Density-based clustering algorithms are independent of earlier knowledge of number of cluster. K-mean algorithm may be useful in situations where the number of cluster should be determined easily before the start of the algorithm [1].In this paper, the analysis of K-mean and Grid density clustering algorithm is done, and also the explain the features in which they are apart from each other.

2. K-Mean

The k-means algorithm partitions the dataset into 'k' subsets. In this, a cluster is represented by its centroid, which is an average point also called mean of points within a cluster. This algorithm works resourcefully only with numerical attributes. And it can be harmfully affected with the single outlier. It is the most accepted clustering algorithm that is used in scientific and industrial applications. It is a way of cluster analysis which aims to partition "n" observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

The basic algorithm is very simple as: [5]

1. Select "k" points as initial centroids.
2. Repeat.
3. Form "k" clusters by assigning each point to its closest centriod.
4. Recompute the centroid of each cluster until centroid does not change.

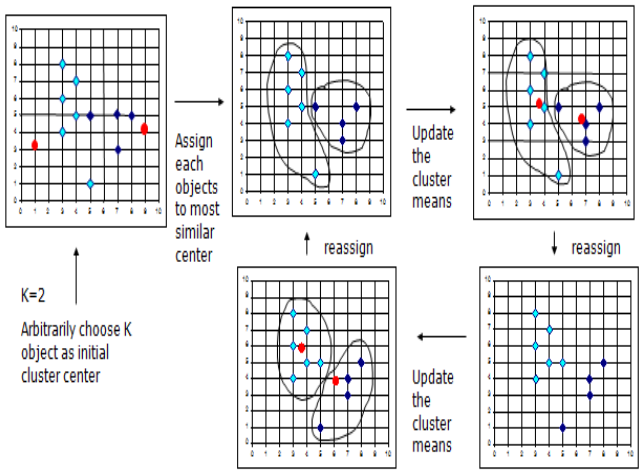

Figure 3: Example of K-mean

As in the above figure the working of the k-mean algorithm is shown. Initially the dataset is present with "n" objects. According to first step of the algorithm, it arbitrarily chooses the value of "k" i.e. 2 as initial cluster center. After that, finding the objects that has the most similar center and put that objects into one cluster. Then update the cluster mean value and reassign the objects according to the most similar center and again update the cluster mean value. This process is repeated again and again until all the objects are put into cluster.

K-mean algorithm has the limitation that it cannot handle noise. It is based on the number of objects present and only applicable when mean is defined, cannot apply on categorical data. It is also not able to find the non-convex shape clusters.

3. Grid Density

It determines dense grids based on densities of their neighbors. Grid density clustering algorithm is able to handle different shaped clusters in multi-density environments. Grid density is defined as number of points mapped to one grid. The density of a grid is defined by the number of the data points in the grid and if it is higher than density threshold, it is considered as a dense unit. A grid is called *sporadic* when its density is less than the input argument MinPts. [6]. If the *Eps* neighborhood of the grid contains at least a minimum number, *MinPts*, of objects, then the grid is called a core grid. If it is within the *Eps* neighborhood of a core grid, but it is not a core grid, then the grid is called a border grid. A border grid may fall into neighborhood of one or more core grid. For a given grid, if the *Eps* neighborhood of the grid contains less than a minimum number, *MinPts*, of objects, and it is not within the *Eps* neighborhood of any core grids, then the grid is called a noise grid. [8]. In order to obtain superior results it is truly essential to set the parameters accurately. In this, each data record in data stream maps to a grid and grids are clustered based on their density.

The basic 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 for each cell
3. Sorting of the cells according to their densities.
4. Identifying cluster centers.
5. Traversal of neighbor cells.

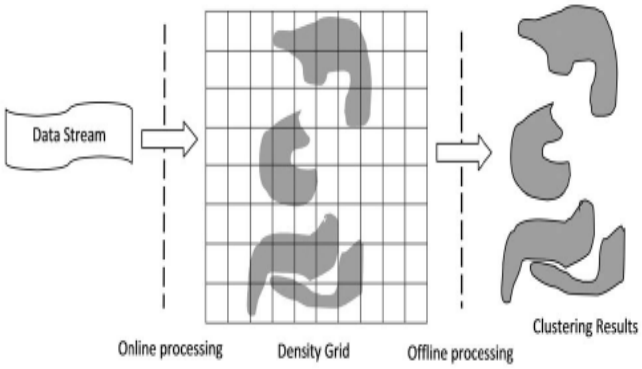

Figure 4 - Density-Grid based clustering Framework [7]

It uses a two-phase scheme, which consists of an online component that processes raw data stream and produces review statistics and an offline component that uses the review data to create clusters. The grid density algorithms are able to handle the noise. In this, the numbers of clusters are not in known in advance. It is also used for spatial database. It can handle the arbitrary shaped clusters.

The table1 shows the comparison of k-mean with grid density algorithms. The table 2 shows the various density and grid density algorithms also show their time complexity and the limitations of the respective algorithm. From the comparison of the two tables, it is clear that grid density algorithm is better than other clustering algorithms.

Table 1. Comparison of K-mean and Grid density algorithms

Parameters	**K-mean**	**Grid density**
Noise	No	Yes
Criteria used	Partition the dataset into "k" clusters	Partition the dataset based on density
No. of clusters known in advance	Yes	No
Used for spatial database	No	Yes
Used for high dimensional	No	Yes
Shape of clusters	Not suitable for non-convex shape clusters	Find the arbitrary shaped clusters
Type of data	Handle numeric data, not handle categorical data.	Not such type of any limitation

4. Related Work in Grid Density and Density

Clustering is a problematical task in Data Mining and Knowledge Discovery. The nearly everyone well-known algorithm in this family is DDBSCAN which takes two input arguments, epsilon and MinPts. Dense regions in dataset are based on these arguments. A region is dense in the order of a exact point if in its epsilon neighborhood radius there are at least MinPts points and these points are called core. If two core points are neighbors of each other, DBSCAN merges these two points. Time complexity of algorithm is $O(n^2)$.

OPTICS is another density based clustering algorithms. It tries to overcome the problem of DBSCAN algorithms by proposing a supplement ordering for points in dataset. This visualized illustration of objects helps the user to choose suitable epsilon value for DBSCAN to achieve a proper result. But this algorithm still has trouble in datasets with overlapped clusters. The reason behind this is that OPTICS is structurally similar to the DBSCAN, and the time complexity of OPTICS is also similar to DBSCAN.

VDBSCAN has the idea of the algorithm is choosing number of epsilon value rather than one epsilon The time complexity of algorithm is O(time complexity of DBSCAN * T), in which "T" is the number of iterations of algorithm.

LDBSCAN suggested using the concepts of local outlier factor (LOF) and local reach ability distance (LRD). It powerfully manipulates the output of clustering but the algorithm has no direction to select correct values.

GMDBSCAN is a grid-density based algorithm to identify clusters. It divides the data space into a number of grids. The computational and time complexity both are high. It works on local density value.

MSDBSCAN settlement from the concept of local core distance(*lcd*). By changing the core point into local core distance, it enables algorithm to find clusters with different shapes in multi density environments. The time complexity of the algorithm is still the drawback of the algorithm like DBSCAN.

GDCLU is the Grid Density Clustering Algorithm which concentrates on time complexity. As it is already discussed in the section2

Table 2: Comparison of density and grid density algorithms

Algorithm	Review
DBSCAN (Density Based Spatial Clustering Algorithm in Noise)	Time complexity is $O(n)^2$ and not suitable with different densities.
OPTICS (Ordering Points To Identify The Clustering Structure)	Time complexity is also $O(n)^2$ and has trouble of overlapped clusters.
VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise)	Time complexity is O (time complexity of DBSCAN* T), where T is the number of iterations of algorithm. And choose a number of epsilon value rather than one value.
LDBSCAN (A local-density based spatial clustering algorithm with noise)	It powerfully manipulates the output of clustering but the algorithm has no direction to select correct values.
GMDBSCAN (Multi-Density DBSCAN Cluster Based on Grid)	The computational and time complexity both are high. It works on local density value.
P-DBSCAN (Photo- Density Based Spatial Clustering Algorithm in Noise)	It concentrates on clustering spatial data. It changes the core point into photo.
MSDBSCAN (Multi Density Scale Independent Clustering Algorithm Based on DBSCAN)	The time complexity of the algorithm is still the drawback of this algorithm.
GDCLU (Grid Density Clustering algorithm)	It generates major clusters by merging dense grids. The time complexity of the algorithm is better than others.

4. CONCLUSION AND FUTURE SCOPE

Grid density takes the advantage of the density and the grid algorithms. Grid density is suitable for handling noise and outliers. 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, our aim is to implement the grid density clustering algorithm for analyze and increase the speed, efficiency and accuracy of the dataset.

ACKNOWLEDGEMENT

I express my sincere gratitude to my guide Ms. Navneet Kaur, for her valuable guidance and advice. Also 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] Mariam Rehman, Syed Atif Mehdi. **Comparison Of Density-Based Clustering Algorithms,** research work, Lahore College for Women University Lahore, Pakistan.

[2]Pasi Fränti. **Number of clusters** (validation of clustering), Speech and Image Processing Unit School of Computing University of Eastern Finland.

[3] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. **Automatic subspace clustering of high dimensional data for data mining applications.** SIGMOD'98.

[4] K. Mumtaz1 and Dr. K. Duraiswamy. **A Novel Density based improved k-means Clustering Algorithm – Dbkmeans.** (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 02, pp. 213-218, 2010.

[5] Pradeep Rai, Shubha Singh. **A Survey of Clustering Techniques,** International Journal of Computer Applications (0975 – 8887), Volume 7– No.12, October 2010.

[6] Gholamreza Esfandani, Mohsen Sayyadi. **GDCLU: a new Grid-Density based CLUstring algorithm,** ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp 102-107, 2012.

[7] Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani, Saeed Reza Aghabozorgi Sahaf Yazdi**. A Study of Density-Grid based Clustering Algorithms on Data Streams,** Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011.

[8] Zheng Hua, Wang Zhenxing, Zhang Liancheng, Wang Qian. **Clustering Algorithm Based on Characteristics of Density Distribution,** National Digital Switching System Engineering & Technological R&D Center, pp431-435, 2010.

[9] D. Gibson, J. Kleinberg, and P. Raghavan. **Clustering categorical data: An approach based on dynamic systems.** VLDB'98.

[10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. **A density-based algorithm for discovering clusters in large spatial databases**. KDD'96.

[11] B. Borah and D.K. Bhattacharyya. **An Improved Sampling-Based DBSCAN for Large Spatial Databases,** International Conference on Intelligent Sensing and Information, IEEE, pp. 92-96, 2004.

[12] Jiawei Han and Micheline Kamber. **Data Mining: Concepts and Techniques**, Morgan Kaufmann Publishers, pp. 383-463, 2006.

[13] Mihael Ankerst, Markus M.Breunig and Hans-Peter Kriegel. **OPTIC:Ordering points to identify the clustering structure,** ACM SIGMOD International Conference on Management of Data, pp. 49-60, 1999.

[14] D. Zhou, N. Wu, P. Liue. **VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise**, International Conference on Service Systems and Service Management, pp. 1-4, 2007.

[15] M. Yufang, Z. Yan, W. Ping, C. Xiaoyun. **GMDBSCAN: Multi- Density DBSCAN Cluster Based on Grid**, IEEE International Conference on e-Business Engineering, pp. 780-783, 2008.

[16] G. Esfandani, H. Abolhassani. **MSDBSCAN: Multi Density Scale Independent Clustering Algorithm Based on DBSCAN**, Advanced Data Mining and Application (ADMA), LNCS, vol. 6440, pp. 202-213, 2010.

Amandeep Kaur Mann pursuing M.Tech in Computer Science & Technology in RIMT institutes of Engineering & Technology, Mandi Gobindgarh belongs to Punjab Technical University. And doing the research work in data mining on grid density algorithm used in Clustering

Navneet Kaur works as Assistant Professor in Computer Science & Technology in RIMT institutes of Engineering & Technology; Mandi Gobindgarh belongs to Punjab Technical University. She guides the M.Tech students in their research work in Data Mining field.