Survey On Different Grid Based Clustering Algorithms Of Data Mining

Nidhi Ashwin Shah[1], Rajkumar Paul[2]

[1]*Computer Science & Engineering Department, Vedica Institute Of Technology, Bhopal, India*
[2]*Computer Science & Engineering Department, Vedica Institute Of Technology, Bhopal, India*

ABSTRACT

Data mining is the method of finding the useful information in huge data repositories and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. Clustering is a fundamental unsupervised data mining technique which is loosely defined as a process of arranging data objects into clusters based on similarity measures. These days the clustering plays a major role in every day-to-day application. Grid-based methods are highly popular compared to the other conventional models due to their computational efficiency but to find optimal grid size is a key feature in grid-based clustering algorithm. There exists some algorithm in that they achieve optimal grid size but in real life data can be dense or sparse.

Keyword: - *Data mining,KDD;Clustering,grid, GRPDBSCAN, GDILC,GGCA,OPT-GRID(S),STING,CLIQUE*

1. 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. Data mining is one of the steps of "Knowledge Discovery from Dataset" process. This process discover interesting pattern from large dataset by performing data cleaning, integration, selection, mining, pattern evaluation and knowledge presentation. The overall goal of data mining is to extract information from a dataset and transform it into an understandable structure for further use. However there are many problems exist in mining data in large datasets such as data redundancy, value of attribute is not specific; data is not complete.

Clustering is an unsupervised classification of patterns (data items, feature vectors or observation) into groups (clusters) which had been address in different contexts by many researchers in different domains across the globe. 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. A good clustering method will produce high quality clusters in which (1) the intra-class (that is, intra-cluster) similarity is high. (2) The inter-class similarity is low.

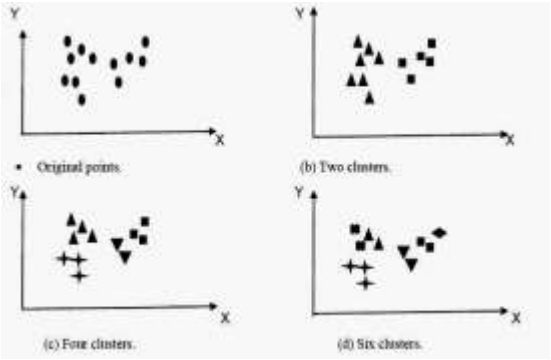

Fig 1 : Different types of clusters[9]

There are various applications of clustering such as Economic Science (especially market research), Document classification, Cluster Weblog data to discover groups of similar access patterns, Pattern Recognition, Create thematic maps in GIS by clustering feature spaces, Image Processing, Information Retrieval, Text Mining.

2. GRID BASED CLUSTERING

The grid-clustering algorithm is the most important type in the hierarchical clustering algorithm. The grid-based clustering approach considers cells rather than data points. This is because of its nature grid-based clustering algorithms are generally more computationally efficient among all types of clustering algorithms. In fact, most of the grid-clustering algorithms achieve a time complexity of where n is the number of data objects. It allows all clustering operations to perform in a gridded data space.

Grid-based clustering where the data space is quantized into finite number of cells which form the grid structure and perform clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data.

Grid-based methods are highly popular compared to the other conventional models due to their computational efficiency. The main variation between grid-based and other clustering methods is as follows. In grid-based clustering all the clustering operations are performed on the segmented data space, rather than the original data objects [1]. Then any topological neighbor search is used to group the points of the closer grids. The grid-based clustering methods face the following challenges. First, it has to determine an appropriate size of the grid structure. If the grid size is too large, two or more clusters may be merged into single one. When the grid size is very small, a cluster may be divided into several sub-clusters. Therefore, finding the suitable size of grid is a challenging issue in grid clustering methods. The second problem is with the data of clusters with variable densities and arbitrary shapes in case of which a global density threshold cannot result the clusters with less densities. This is known as the problem of locality of cluster. The third one is the selection of merging condition to form efficient clusters. Considering these issues, various grid based algorithms have been proposed. Thus these algorithms have a fast processing time, because they go through the data set once to compute the statistical values for the grids and the performance of clustering depends only on the size of the grids which is usually much less than the data objects.

Grid-based methods are widely used in today's various fields applications such as Pattern Recognition, Spatial Data Analysis, Image Processing, WWW(World Wide Web) and others.

3. INTRODUCTION OF VARIOUS GRID BASED ALGORITHMS

3.1 A New Shifting Grid Clustering Algorithm [2]

The algorithm is based on shifting the grid. The algorithm is a non-parametric type means which does not require users inputting parameters. It divides each dimension of the data space into certain intervals to form a grid structure in the data space, shifting of the whole grid structure is introduced to obtain a more descriptive density profile based on the concept of sliding window. As a result, we are able to improve the accuracy of the results. This algorithm is provides effective and efficient clustering because it clusters data in a way of cell rather than in points. It exhibits tremendous performance to deal with arbitrary shaped cluster. The algorithm does not always suffer from the problem of memory limitation when handling large data set. The algorithm has disadvantage is that it requires the additional computational efforts.

- **Advantages:**

(1) Its computational time is independent to the number of data points.

(2) It shows good performance to deal with arbitrary shaped cluster

(3) It is a non-parametric algorithm.

(4) The algorithm does not always suffer from the problem of memory limitation When handling large data set.

3.2 Grid-Based DBSCAN Algorithm With Referential Parameters [3]

The algorithm which is combination of the grid partition technique and multi-density based clustering algorithm. The key issue is to be how to deal with the data changes and how assure the validity of data class' association rules. The algorithm solved this problem .According to the character of data mutations in dynamic data test, and the association between grid partition technique and multi-density base clustering algorithm: DBSCAN is used. The algorithm chooses the Eps and Minpts parameters automatically and differentiates between noises and discovery clusters of arbitrary shapes including data changes, so they were more objective (independent). The algorithm has improved its efficiency. Finally, the algorithm makes answer more precise and the result of this improved algorithm is robust.

- **Advantages:**

(1) It deals with massive data changes and noise.

(2) It can find clusters of arbitrary shape and remove noise.

(3) It makes the answer more precise and it is more robust.

3.4 GDILC : A Grid-Based Density-Isoline Clustering Algorithm [4]

The idea of GDILC is that the density-isoline figure depicts the distribution of data sample very well. Here, assumes that the entire data samples are normalized. The algorithm starts from the density-isoline figure of data samples to find the densely populated regions, which meets the requirements of the clustering or which are the clusters that we hope to find. In that first, a grid-based method is working to calculate the density of each data sample and chosen a proper density threshold from density-isoline figure. Then, those clusters bounded by the chosen isoline are combined. GDILC can calculate automatically the distance threshold and the density threshold according to the size and distribution of a given data set. So it is non-supervising clustering algorithm because it require no human iteration. The advantage of these algorithm is the high speed and accuracy & mainly removing outlier and finding the clusters of various shapes.

- **Advantages:**

(1) It has the cluster of arbitrary shapes and locate the outliers.

(2) It has linear time complexity.

3.5 A General Grid-Clustering Approach [5]

A general grid-clustering approach (GGCA) under a common assumption about hierarchical clustering. The key features of the GGCA include: (1) It is the combination of the divisible and the agglomerative clustering algorithms into a unifying generative framework; (2) the purpose of key input parameters: an optimal grid size for the first time; and (3) the application of a two-phase merging process to combined all data objects. The GGCA is a non-parametric algorithm in which it does not require users to input parameters. With the partitioning index and the depth index, the algorithm solves two critical problems of conventional (predictable) grid-clustering algorithms: (i) grid size and (ii) merging condition.

The characteristics of GGCA are: (1) GGCA consists of a set of local high-density data objects and, therefore, the problem in the existing clustering algorithms is that a global density threshold is unable to identify all clusters in a given density-diverse dataset. Also, the GGCA does not reject those low-density clusters, while many other existing clustering approaches often may not be able to find them. (2) A good clustering method should have the ability to accept noisy data and outliers in the dataset. Using a two-phase merging process enhancements the robustness of the clustering results by the GGCA. Since the boundary data objects of each cluster are often located in the overlapped area of diverse clusters, these leads to incorrect clustering results in the existing clustering approaches. However, these boundary objects cannot be contained in any core grid, and thus do not affect the clustering results, so the risk of incorrect clustering results by a clustering algorithm decreases. GGCA gives excellent performance in dealing with not well-separated and shape-diverse clusters.

- **Advantages:**

(1) It is parameter-free algorithm.

(2) It has the ability to bear noisy data and outlier in the dataset

3.6 OPT-GRID(S) [12]

The grid-based clustering finds the optimal gird-size using the boundaries of the clusters. Here, In dataset D there are n data points and m dimensional space. Initially a single grid which is used to represent all the given n points. This grid are taken as the min. and max. attribute value in each dimension. Then single grid is partitioned into two equal volume grids. So, all the data points are distributed to these two grids. After each round of partitioning of gird it is necessary to check the presence of the new cluster. In the next round of partitioning, the two grids are partitioned into four equal volume grids in another chosen dimension. In this way all the grids are bisected and partitioning processes is continue until the optimal grid structure is generated. The boundary grids are used to find the optimal grid structure. Each cluster is surrounded by the boundary grids. The volume of the cluster decreases as the partitioning process continues and as the same time the number of surrounding boundary grids increases. The problem of outlier is solved with local outlier factor (LOF).

- **Advantages:**

(1) It is parameter-free algorithm.

3.7 STING(Statistical Information Grid Approach) [8]

The spatial area is divided into rectangular cells, There are several levels of cells corresponding to different levels of resolution. Used a top-down approach to answer spatial data queries. Start from a pre-selected layer—typically with a small number of cells. From the pre-selected layer until you reach the bottom layer do, For each cell in the current level compute the confidence interval indicating a cell's relevance to a given query; If it is relevant, include the cell in a cluster If it irrelevant, remove cell from further consideration otherwise, look for relevant cells

at the next lower layer, Combine relevant cells into relevant regions (based on grid-neighborhood) and return the so obtained clusters as your answers.

- **Advantages:**

(1) Query-independent, easy to parallelize, incremental update.

(2) Execution time, *O(K),* where *K* is the number of grid cells at the lowest level.

3.8 CLIQUE(Clustering In Quest)[11]

Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space CLIQUE can be considered as both density-based and grid-based, It partitions each dimension into the same number of equal length interval, It partitions an m-dimensional data space into non-overlapping rectangular units, A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter. A cluster is a maximal set of connected dense units within a subspace.

- **Advantages:**

(1) It *automatically* finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces.

(2) It is *insensitive* to the order of records in input and does not presume some canonical data distribution.

(3) It scales *linearly* with the size of input and has good scalability.

4. CONCLUSIONS

Grid based clustering is efficient and good technique to make cluster of data. Iteratively grids are created and according to that clusters will created. This paper gives detail study of the various grid based algorithms such as GRPDBSCAN, GDILC, GGCA, OPT-GRID(S), STING, CLIQUE based on the different parameters which gives different arbitrary shapes. To develop an algorithm that can find optimal grid size in any type of dataset in dense or sparse with appropriate accuracy or maintaining the accuracy with less time is challenging task in a grid clustering.

5. REFERENCES

[1].Damodar Reddy Edla and Prasanta K. Jana "A Grid Clustering Algorithm Using Cluster Boundaries" IEEE World Congress on Information and Communication Technologies 2012

[2]. E. W. M. Ma and T. W. S. Chow, "A new shifting grid clustering algorithm," Pattern Recognition, vol. 37, pp. 503-514, 2004.

[3]. H. Darong and W. Peng, "Grid-based DBSCAN Algorithm with Referential Parameters," Proc. International Conference on Applied Physics and Industrial Engineering (ICAPIE-2012), Physics Procedia, vol. 24(B), pp. 1166-1170, 2012.

[4].Y.Zhao and J. Song, GDILC: A Grid-based Density-Isoline Clustering Algorithm," Proc. International Conferences on Info-tech and Info-net (ICII-2001), vol. 3, pp. 140-145, October 29-November 1, 2001.

[5]. N. Chen, A. Chen and L. Zhou, "An incremental grid density-based clustering algorithm," Journal of Software, vol. 13, no. 1, pp. 1-7, 2002.

[6]. 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.

[7]. Oded Maimon, Lior Rokach, "**DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK",** Springer Science+Business Media.Inc, pp.321-352, 2005.

[8]. Jiawei Han**,** Micheline Kamber, **"Data Mining:Concepts and Techniques Second Edition",**University of Illinois at Urbana-Champaign, 2006 by Elsevier Inc.

[9]. Amandeep Kaur Mann, Navneet Kaur, "Grid Density Clustering Algorithm", International Journal of Innovative Research in Science,Engineering and Technology, Vol. 2, Issue 10, October 2013

[10]. Pang-Ning Tan Michael Steinbach Vipin Kumar, **"Introduction to Data Mining"**, Copyright c 2006 Pearson Addison-Wesley

[11]. MR ILANGO,Dr V MOHAN, "A Survey of Grid Based Clustering Algorithms", Ilango et. al. / International Journal of Engineering Science and Technology Vol. 2(8), 2010, 3441-3446

[12]. Monali Parikh,Tanvi Varma, " IOG -An Improved Approach to Find Optimal Grid Size Using Grid Clustering Algorithm", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IV (May-Jun. 2014), PP 114-118