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
www.iosrjournals.org

IOG - An Improved Approach to Find Optimal Grid Size Using Grid Clustering Algorithm

Monali Parikh[1], Asst. Prof. Tanvi Varma[2]
[1](*Computer Science Engineering Department, Gujarat Technological University, India)*
[2](*Computer Science Engineering Department, Parul Institute of Technology, India)*

Abstract: *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. In grid-based clustering, all the clustering operations are performed on the segmented data space, rather than the original data objects. 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 exist some algorithm in that they achieve optimal grid size but in real life data can be dense or sparse. So, in these research 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.*
Keywords: *Data mining; Clustering; Grid; k-nn method*

I. Introduction

Data mining is the extraction of hidden, predictive information patterns from large databases**.** Data mining definition can be described as a process of analyzing and then re-arranging the patterns of the data and finding co-relations in them in such a way that it goes in the benefit of the business overall. Data mining is a combination of three main factors: Data, Information and knowledge. Data are the most elementary description of the things, events or the activity and transactions. Information is organized data which have some valuable meaning or some useful data. Knowledge is a concept of understanding information based on the recognized pattern or algorithms that provide the information. Data Mining is a technique of finding valuable knowledge from the large amount of dataset [1] [2]. Data mining is also called knowledge discovery from the huge amount of data.

A cluster is a subset of objects which are "similar". Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. A good clustering method will generate high value clusters in which (1) the intra-class (that is, intra-cluster) relationship is high. (2) The inter-class (that is, inter-cluster) relationship is low [3]. The main clustering methods are partitioning clustering, hierarchical clustering, density-based clustering, model-based clustering and grid 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 $O(n)$where n is the number of data objects. It allows all clustering operations to perform in a gridded data space. 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. Then any topological neighbor search is used to group the points of the closer grids [4]. The grid-based clustering uses the multi-resolution grid data structure. It is non-parametric means it does not require users to input parameter. Gird-based method could be ordinary choice for data stream in which the vast data streams map to fixed grid cells. The summary information for data streams is controlled in the grid cells. The example of grid-based clustering are STING (a STatistical INformation Grid approach), CLIQUE which is applied on high dimensional data and wavecluster.

The grid-based clustering method has the following advantages. (1) Shapes are limited to union of grid-cells. (2) It has fast Processing time in terms of it does not calculate distance and it is easy to define which clusters are bordering (neighboring). Also, clustering is performed on summaries and not individual objects. The grid-based clustering methods face the following challenges [4]. 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.

Basic Steps of Grid-based Clustering Algorithms are as follows [5]:

(1) Define a set of grid-cells.

(2) Assign objects to the appropriate grid cell and compute the density of each cell.

(3) Eliminate cells, whose density is below a certain threshold t.

(4) Form clusters from adjacent (neighboring) groups of dense cells.

The remaining paper is organized as follows: section 2 background and related work, the section 3 contains the proposed method details, and section 4 provides result analysis and comparisons. Finally, in the section 5 some conclusions are drawn.

II. Background And Related Work

Here, various algorithms are presented which are related to grid-based clustering. Grid-based DBSCAN Algorithm with Referential Parameters (GRPDBSCAN) [6] which is combination of the grid partition technique and multi-density based clustering algorithm. The algorithm solves how to deal with the data changes and how assure the validity of data class' association rules. It can find clusters of arbitrary shape and remove noise. It makes the answer more precise and it is more robust. A general grid-clustering approach (GGCA) [7] which is under a common assumption about hierarchical clustering. The GGCA is a non-parametric algorithm in which it does not require users to input parameters. GGCA gives excellent performance in dealing with not well-separated and shape-diverse clusters. A new shifting grid clustering algorithm [8] which is based on shifting the grid. It divides each dimension of the data space into certain intervals to form a grid structure in the data space. Its computational time is independent to the number of data points. The algorithm does not always suffer from the problem of memory limitation when handling large data set. A Grid-based Density-Isoline Clustering Algorithm (GDILC) [9] which 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 requires no human iteration. The advantage of these algorithm is the high speed and accuracy & mainly removing outlier and finding the clusters of various shapes. It has linear time complexity.

Sr. no	Algorithm Name	Technique Used	Input parameter	Arbitrary shape	Noise
1	GRPDBSCAN	Grid + Density	Eps and Minpts	Yes	Yes
2	GGCA	Hierarchical Grid + Density	Non-parametric (parameter-free)	No	Yes
3	Shifting grid clustering algorithm	Grid + Density	Non-parametric (parameter-free)	Yes	No
4	GDILC	Grid + Density	distance threshold RT and density threshold DT	Yes	Yes

Table 1. Comparison of Grid based Algorithms

III. Proposed Method

The proposed work is to find optimal grid size 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. So, 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.

The proposed algorithm works in same manner except stopping criteria. It uses the k-nn for stopping criteria. It generate cluster using the OPTGRID's connected () function. For each point of the cluster find the k-nn of it. If the k-nn are only from the same cluster than and only than stopping condition is full filled, otherwise continue the partitioning process and cluster generation.

With generation of clusters outlier detection is also required. If there exist outlier in the cluster it changes the characteristic of the cluster. The proposed approach uses k-nn for this problem. Thus if all the k-nn of the point present in the same cluster than and only than cluster does not contain any outlier, otherwise there exist some outlier and we need to continue the partitioning process and clustering.

Algorithm Steps:

(1) Call G(n,m) which is function to find the initial grid structure of the given set S of n points with dimension m.

(2) This initial grid is taken as the minimum and maximum attribute value in each dimension. Then the initial grid G(n,m)=[min(1),max(1)]*[min(2),max(2)]*….*[min(m),max(m)].
(3) Now call partition the grid G(n,m) into two equal volume of grids. So, the initial grid G(n,m) is partitioned into G1(n1,m) and G2(n2,m) in a uniformly selected dimension m.
(4) Data point of G(n,m) are distributed to this two grids (G1(n1,m) and G2(n2,m)) which have non-empty and empty grids.
(5) After each round of partitioning of gird it is necessary to check the presence of the new cluster.
(6) In the next round of partitioning, the two grids G1(n1;m) and G2(n2;m) are partitioned into four equal volume grids G1,1(n3;m), G1,2(n4;m); and G2,1(n5;m), G2,2(n6;m) in another chosen dimension.
(7) In this way all the grids are bisected and partitioning processes is continue until the optimal grid structure is generated.
(8) Call Connected() to produce the clusters (say, C1,C2,…,Cl for some l) by grouping points in the grids which are connected by the common vertices.
   (a) Start with any random grid G.
   (b) **If** G is non-visited, **then** mark it as visited and add all the point of G in Cl
      i.e., Cl ← Cl ∪ {points in the grid G}; **Else Go to** Step (a).
   (c) Find the non-visited grids Gr (for some r) shared by G with any common vertex.
   (d) Add the points of all Gr to the cluster Cl and Gr is visited.
      i.e., Cl ←Cl ∪ {points in the grids Gr};
   (e) Repeat the Steps (c) and (d) for all the grids Gr until no new grid is identified.
   (f) **If** all the cluster point size= total number of data (Here, Cj=n) **then Return** (C1, C2,…, Cj)
      Else Go to step (a) to restart the process
(9) Find the number of boundary grids of all the clusters C1, C2,…, Cl
(10) **If** boundary grids of any cluster has outlier **then Go to** step 11
   Else Go to Step 3 to call the grid function for all the grids.
(11) Call k-nn count
(12) **If** k-nn count is less than 60% **then** continue the partitioning process for the same cluster
   Else Go to Step 13
(13) Output the generated clusters C1, C2,…, Ck with respect to the grid size.

IV. Result Analysis

To run grid clustering algorithm, java platform is used and Eclipse Juno is used as development tool. At the end of execution, it will be observed time for process. Various data sets are used for taking results. Grid based clustering is used for partition method. Initially partition will be created and based on that cluster will be generated.

In the proposed system, the Heart dataset, Iris dataset and Wine dataset are used to perform grid clustering algorithm and all are taken from UCI Machine Learning Repository which is having the extension .txt [10]

Data	No. of Attributes	Instances
Heart	13	270
Iris	4	150
Wine	13	178

Table 2. Dataset Description

The algorithm is used the normalized information gain (NIG) to calculate the quality of the clusters. Normalized Information Gain is a measure of the quality of clusters which is developed based on the information gain. NIG is expressed in terms of total entropy and weighted entropy. The Total Entropy (ENTotal) which is nothing but the average information per point in the dataset measured in bits. The weighted entropy (wEN) which is nothing but the average information per point in each cluster. The smaller values of NIG indicate more quality of the clusters.

The presented execution result of OPTGRID and IOG with specific dataset and parameter are shown. Also, the execution time vs. algorithm with different dataset and value of NIG with different data set are shown.

Dataset	No. of Attributes	Instances	No. of Clusters	NIG	
				OPTGRID	IOG
Iris	4	150	3	0.98	0.98
Wine	13	178	3	0.99	0.99
Heart	13	270	2	0.95	0.95

Table 3. Dataset VS NIG (Clusters and NIG Value)

The above table shows the presented result of OPTGRID algorithm and IOG algorithm. The real data sets like iris, wine and heart is used. As shown in table NIG (Normalized Information Gain) in OPTGRID algorithm and IOG algorithm are same. So it is clearly observed that IOG algorithm is also giving similar result according to OPTGRID algorithm.

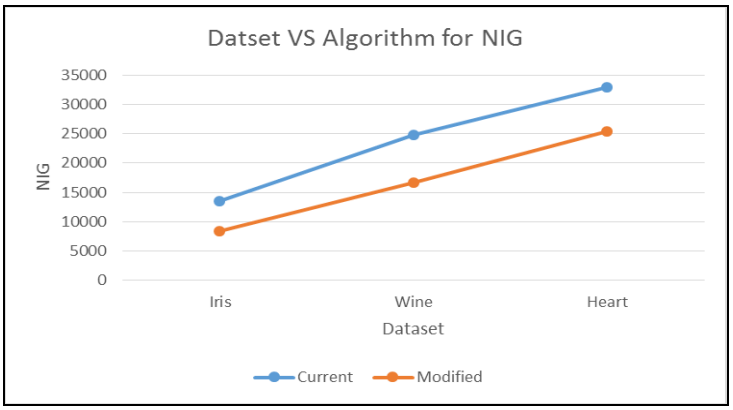

Figure 1. Dataset VS Algorithm for Value of NIG

Dataset	Time (ms)	
	OPTGRID	IOG
Iris	13584	8365
Wine	24833	16745
Heart	32865	25379

Table 4. Dataset VS NIG (Time)

As shown in table, by comparing the IOG algorithm with OPTGRID algorithm with using time as parameter. It can be observed that execution time is less while using IOG algorithm. So, using this result it can be clearly analyzed that IOG algorithm is more efficient than OPTGRID algorithm. Graphical represent of result is available below.

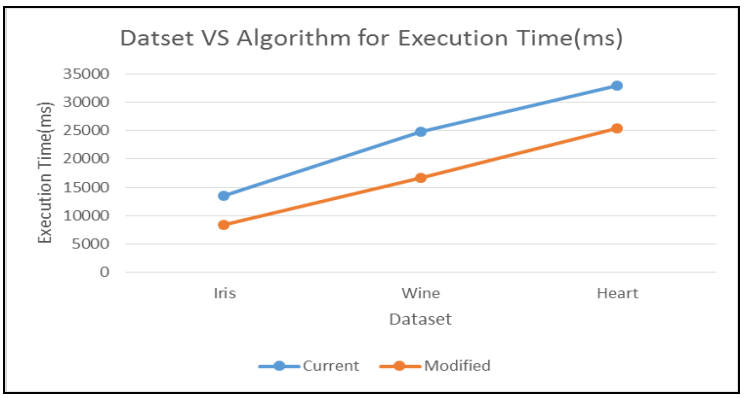

Figure 2. Dataset VS Algorithm for Execution Time (ms)

V. Conclusion

The Experimental results show that the IOG approach consumes less execution time with same number of clusters as compared to OPTGRID approach. OPTGRID takes more time to calculate the LOF whereas IOG approach uses less time because it uses the k-nn approach for stopping criteria. In the k-nn method, the distance is calculated only once, while in the OPTGRID the calculation for the LOF increase the overhead. Even the k-nn will also be able to remove the outliers same as the LOF but will optimize the grid partitioning.

Acknowledgements

I would like to thank Assistant Professor Ms. Tanvi Varma of Parul Institute of Technology, Baroda, India for his constant guidance and support in this research.

References

[1] Han, P.N., Kamber, M.: Data Mining: Concepts and Techniques,2nd (2006).
[2] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining (2006).
[3] http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf
[4] Damodar Reddy Edla and Prasanta K. Jana "A Grid Clustering Algorithm Using Cluster Boundaries" IEEE World Congress on Information and Communication Technologies 2012
[5] https://www.google.co.in/?gfe_rd=cr&ei=k719U5PTMc_N8gfYvoGwBg#q=dm_clustering2.ppt
[6] 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
[7] 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.
[8] E. W. M. Ma and T. W. S. Chow, "A new shifting grid clustering algorithm," Pattern Recognition, vol. 37, pp. 503-514, 2004.
[9] 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.
[10] UCI Machine Learning Repository ,http://archive.ics.uci.edu/ml/datasets.html