INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY
© VISHWASHANTI MULTIPURPOSE SOCIETY (Global Peace Multipurpose Society) R. No. MH-659/13(N)
www.vmsindia.org

Grid-Based Clustering: An Overview

P. B. Dhumane[1] and S. R. Pande[2]
[1]Department of Comp. Science, Sardar Patel Mahavidyalaya, Chandrapur
[2]Department of Comp. Science, SSES Science College, Nagpur

Abstract :
Grid based clustering algorithms are efficient in mining large multidimensional data sets[1]. These algorithms partition the data space into a finite number of cells to form a grid structure and then form clusters from the cells in the grid structure. The efficiency of grid based clustering algorithms comes from how data points are grouped into cells and clustered collectively rather than individually. This results in drastic time complexity improvements because often data is grouped into far few cells then there are data points[1]. The general approach of these algorithms is to divide the data space into grid data structures, summarize each cell in the grids by a statistic such as density, and then cluster the grid cells[1].

Introduction :
Grid based clustering algorithms are efficient in mining large multidimensional data sets. These algorithms partition the data space into a finite number of cells to form a grid structure and then form clusters from the cells in the grid structure. Clusters correspond to regions that are more dense in data points than their surroundings[1]. Grids were initially proposed by Warnekar and Krishna[2] to organise the feature space, e.g. in GRIDCLUS[3] , and increased in popularity after STING[4], CLIQUE[5] were introduced. The great advantage of grid based clustering is a significant reduction in time complexity, especially for very large data sets. Rather than clustering represented by cells[1]. In most applications since the number of cells is significantly smaller than the number of data points. Grid based clustering algorithms typically involve the following five steps[6][7]:

1. Creating the grid structure, i.e. partitioning 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 neighbour cells.

A. GRIDCLUS

Schikuta introduced the first GRID-based hierarchical CLUStering algorithm called GRIDCLUS[3]. The algorithm partitions the data space into a grid structure composed of disjoint d-dimensional hyper rectangles or blocks. Data points are considered points in d-dimensional space and are designated to blocks in the grid structure such that their topological distributions are maintained. Once the data is assigned to blocks, clustering is done by a neighbour search algorithm. The GRIDCLUS algorithm is given in algorithm 1 and the function NEIGHBOR_SEARCH is the recursive procedure given in algorithm 2 [7][3].

Algorithm 1 GRIDCLUS Algorithm

1. Set u = 0, W[] = { }, c[] = { } { initialization}
2. Create the grid structure and calculate the block density indices .
3. Generate a sorted block sequence $B_{1'}, B_{2'}, \ldots, B_{b'}$ and mark all blocks "not active" and "not clustered".
4. While a "not active" block exists do
5. $u = u+1$
6. mark first $B_{1'}, B_{2'}, \ldots, B_{j'}$ with equal density index "active"
7. for each "not clustered" block $B_i = B_{1'}, B_{2'}, \ldots, B_{j'}$ do
8. Create a new cluster set C[u].
9. W[u] = W[u] + 1, C[u,W[u]]= $B_{i'}$.
10. Mark block $B_{i'}$ clustered.
11. NEIGHBOR_SEARCH($B_{i'}$, C[u,W[u]]).
12. end for
13. For each "not active" block B do
14. W[u] = W[u] +1, C[u,W[u]] = B
15. end for
16. Mark all blocks "not clustered"
17. end while

Algorithm 2 NEIGHBOR_SEARCH(B,C) procedure

1. for each "active" and "not clustered" neighbour B' of B do
2. $C = B'$
3. Mark block B' "clustered"
4. NEIGHBOR_SEARCH (B',C)
5. End for

B. BANG

The BANG algorithm introduced by Schikuta and Erhart is an extension of the GRIDCLUS algorithm[8]. It addresses some of the inefficiencies of the GRIDCLUS algorithm in terms of grid structure size, searching for neighbors, and managing blocks by their density. BANG also places data points in blocks and uses

a variant of the grid directory called a BANG structure to maintain blocks. Neighbor search and processing the blocks in decreasing order of density are also used for clustering blocks. Nearness of neighbors is determined by the maximum dimensions shared by a common face between blocks. A binary tree is used to store the grid structure, so that neighbour searching can be done more efficiently. From this tree in the grid directory and the sorted block densities, the dendrogram is calculated. Centers of clusters are still the most highly dense blocks in the clustering phase. The BANG algorithm is summarized in Algorithm 3 [7][8].

Algorithm 3 BANG Clustering Algorithm

1. Partition the feature space into rectangular blocks which contain up to a maximum of pmax data points.
2. Build a binary tree to maintain the populated blocks, in which the partition level corresponds to the node depth in the tree.
3. Calculate the dendrogram in which the density indices of all blocks are calculated and sorted in decreasing order.
4. Starting with the highest density index, all neighbour blocks are determined and classified in the original blocks.
5. Repeat step 4 for the remaining blocks of the dendrogram.

C. STING

Wang et al. [4] proposed a STatisticalINformation Grid based clustering method (STING) to cluster spatial databases and to facilitate region – oriented queries. STING divides the spatial area into rectangular cells and stores the cells in a hierarchical grid structure tree. Each cell is partitioned into 4 child cells at the next level with each child corresponding to a quadrant of the parent cell. A parent cell is the union of its children; the root cell at level 1 corresponds to the whole spatial area. The leaf level cells are of uniform size, determined globally from the average density of objects. For each cell, both attribute dependent and attribute independent parameters of the statistical information are maintained. The algorithm is summarized in Algorithm 4 [7][4]

Algorithm 4 STING clustering algorithm

1. Determine a level at which to begin.
2. For each cell of this level, we calculate the confidence interval of probability that this cell is relevant to the query.
3. From the interval calculated above, label the cell as relevant or not relevant.
4. If this level is the leaf level, go to step 6, otherwise step 5.
5. Go down the hierarchy structure by one level. Go to step 2 for those cells that form the relevant cells of the higher level.
6. If the specification of the query is met, go to step 8; otherwise, go to step 7.
7. Retrieve those data that fall into the relevant cells and do further processing. Return the result that meets that the requirement of the query. Go to step 9.
8. Find the regions of relevant cells. Return those regions that meet the requirement of the query.
9. Stop.

D. **WaveCluster :** Wavelets in Grid-Based Clustering

Sheikholeslami et al. [9]proposed a grid-based and density-based clustering approach that uses wavelet transforms :WaveCluster. This algorithm applies wavelet transforms to the data points and then uses the transformed data to find clusters. A wavelet transform is a signal processing technique that decomposes a signal into different frequency subbands. The insight to using the wavelet transforms is that data points are considered *d*-dimensional signals where *d* is the number of dimensions. This algorithm is given in Algorithm 5 [9]

Algorithm 5 WaveCluster Algorithm

INPUT : Multidimensional data objects feature vectors

OUTPUT : cluster objects

1. First bin the feature space, then assign objects to the units, and compute unit summaries.
2. Apply wavelet transform on the feature space.
3. Find connected components (clusters) in the subbands of transformed feature space, at multiple levels.
4. Assign labels to the units in the connected components.
5. Make the lookup table.
6. Map the objects to the clusters.

Conclusion :

The efficiency of grid based clustering algorithms comes from how data points are grouped into cells and clustered collectively rather than individually[1]. This results in drastic time complexity improvements because often data is grouped into far few cells than there are data points. The general approach of these algorithms is to divide the data space into grid data structures, summarize each cell in the grids by a statistic such as density, and then cluster the grid cell[1].

References

[1] Charu C. Aggarwal and Chandan K. Reddy, Data Clustering : Algorithms and Application, CRC Press.

[2] a. G. K. C.S. Warnekar, "A heuristic clustering algorithm using union of overlapping pettern-cells," in *Pattern Recognition*, 1979.

[3] E. Schikuta, "Grid-Clustering : an efficient hierarchical clustering method for very large data sets," in *13th International Conference on Pattern Recognition-2*, 1996.

[4] J. Y. a. R. R. M. Wei Wang, "STING : A statistical information grid approach to spatial data mining," in *VLDB'97, 23RD International Conference on very Large Data Bases.*, Athens, Greece, 1997.

[5] J. G. D. G. a. P. R. Rakesh Agrawal, "Automatic subspace clustering of highdimensional data for data mining applications," in *ACM SIGMOD International Conference on Management of Data "SIGMOD'98"*, New York, USA, 1998.

[6] P. G. a. A. Borisov, "Using Grid-clustering methods in data classification," in *International Conference on Parallel Computing in Electrical Engineering*, 2002.

[7] C. M. G. G. a. J. Wu, Data Clustering : Theory, Algorithms and Application, SIAM, 2007.

[8] E. S. a. M. Erhart, "The BANG-clustering : Grid-based data analysis," in *2nd International Symosium on Advances in Intelligent Data Analysis, Resoning about Data IDA-97*, Verlag, 1997.

[9] S. C. a. A. Z. Gholamhosein Sheikholeslami, "WaveCluster : A multi-resolution clustering approach for very large spatial databases," in *International Conference on very Large Data Bases VLDB-98*, 1998.