A Survey of Grid Based Clustering Algorithms

MR ILANGO[1]

Dr V MOHAN[2]

[1]Professor, Department of Computer Applications, K L N College of Engineering, Pottapalaym- 630611.

Sivagangai District, Tamilnadu, India

[2] Professor and Head, Department of Mathematics, Thiagarajar College of Engineering, Madurai, Tamilnadu,

India-625015,

Abstract:
Cluster Analysis, an automatic process to find similar objects from a database, is a fundamental operation in data mining. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering techniques have been discussed extensively in Similarity Search, Segmentation, Statistics, Machine Learning, Trend Analysis, Pattern Recognition and Classification [1]. Clustering methods can be classified into i) Partitioning methods ii) Hierarchical methods iii) Density-based methods iv) Grid-based methods v) Model-based methods. Grid based methods quantize the object space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space. The main advantage of Grid based method is its fast processing time which depends on number of cells in each dimension in quantized space. In this research paper, we present some of the grid based methods such as CLIQUE (CLustering In QUEst**) [2]**, STING (STatistical INformation Grid) [3], MAFIA (Merging of Adaptive Intervals Approach to Spatial Data Mining) [4], Wave Cluster [5]and O-CLUSTER (Orthogonal partitioning CLUSTERing) [6], as a survey and also compare their effectiveness in clustering data objects. We also present some of the latest developments in Grid Based methods such as Axis Shifted Grid Clustering Algorithm [7] and Adaptive Mesh Refinement [Wei-Keng Liao etc] [8] to improve the processing time of objects.

Key Words: Clustering, Grids

1.0 INTRODUCTION

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity but are very dissimilar to objects in other clusters. A good clustering algorithm should be able to identity clusters irrespective of their shapes. Other requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitivity to the order of input records, etc.

2.0 CLIQUE (CLUSTERING IN QUEST)

It makes use of concepts of density and grid based methods. In the first step, CLIQUE partitions the n-dimensional data space S into non overlapping rectangular units (grids) [2]. The units are obtained by partitioning every dimension into ξ intervals of equal length. ξ is an input parameter, selectivity of a unit is defined as the total data points contained in it. A unit u is dense if selectivity (u) is greater than γ, where the density threshold γ is another input parameter. A unit in the subspace is the intersection of an interval from each of the K attributes. A cluster is a maximal set of connected dense units. Two K-dimensional units u1, u2 are connected if they have a common face. The dense units are then connected to form clusters. It uses apriori algorithm (bottom up algorithm) to find dense units. The dense units are identified by using a fact that if a K dimension unit $(a_1, b_1)^* (a_2, b_2) \ldots\ldots.(a_k, b_k)$ is dense, then any k-1 dimension unit $(a_1,b_1) * (a_2,b_2)\ldots.(a_{ik-1},b_{ik-1})$ is also dense where (a_i, b_i) is the interval of the unit in the i^{th} dimension.

Given a set of data points and the input parameters ξ and γ CLIQUE is able to find clusters in all subspaces of the original data space and present a minimal description of each cluster in the form of a DNF expression. Steps involved in CLIQUE is i) identification of sub spaces (dense Units) that contain cluster ii) merging of dense units to form cluster & iii) Generation of minimal description for the clusters.

3.0 STING:(A STatistical INformation Grid Approach to spatial Data Mining).

Spatial data mining is the extraction of implicit knowledge, spatial relation and discovery of interesting characteristics and patterns that are not explicitly represented in the databases. (Spatial data mining has wide applications in many fields, including GIS system, image data base exploration, medical imaging etc).

STING is a grid based multi resolution clustering technique in which the spatial area is divided into rectangular cells (using latitude and longitude) and employs a hierarchical structure [3].

There are usually several levels of such rectangular cells corresponding to different levels of resolution. Each cell at a high level is partitioned to form child cells at lower level. A cell in level i corresponds to union of its children at level i + 1. Each cell (except the leaves) has 4 children & each child corresponds to one quadrant of the parent cell.

Statistical information regarding the attributes in each grid cell (such as, mean, Standard Deviation maximum & minimum values) is pre computed and stored.

Statistical parameters of higher level cells can easily be computed from the parameters of lower level cells.
For each cell, there are attribute independent parameters and attribute dependant parameters.

i. Attribute independent parameter: count
ii. Attribute dependant parameters

M- Mean of all values in this cell; S- Standard deviation of all values in this cell
Min – minimum value of the attribute in this cell; Max – minimum value of the attribute in this cell; Distribution – Type of distribution the attribute value follows. Distribution types are normal, uniform exponential & none.
Value of distribution may either be assigned by the user or obtained by hypothesis tests such as X^2 test.

When the data are loaded into the database, the parameters count, m, s, min, max of the bottom level cells are calculated directly from the data.
First, a layer is determined from which the query processing process is to start. This layer may consist of small number of cells. For each cell in this layer we check the relevancy of cell by computing confidence internal. Irrelevant cells are removed and this process is repeated until the bottom layer is reached.

4.0 MAFIA: (Merging of Adaptive Intervals Approach to Spatial Data Mining)

MAFIA proposes adaptive grids for fast subspace clustering and introduces a scalable parallel framework on shared nothing architecture to handle massive data sets [4]. Most of the grid based algorithms uses uniform grids whereas MAFIA uses adaptive grids. MAFIA proposes a technique for adaptive computation of the finite intervals (bins) in each dimension, which are merged to explore clusters in higher dimensions.

Adaptive grid size reduces the computation and improves the clustering quality by concentrating on the portions of the data space which have more points and thus likelihood of having clusters. Performance results show MAFIA is 40 to 50 times faster than CLIQUE, due to the use of adaptive grids. MAFIA introduces parallelism to obtain a highly scalable clustering algorithm for large data sets.

MAFIA proposes an adaptive interval size to partition the dimension depending on the distribution of data in the dimension. Using a histogram constructed by one pass of the data initially, MAFIA determines the minimum number of bins for a dimension. Contiguous bins with similar histogram values are combined to form larger bins. The bins and cells that have low density of data will be pruned limiting the eligible candidate dense units, thereby reducing the computation. Since the boundaries of the bins will also not be rigid, it delineates cluster boundaries more accurately in each dimension. It improves the quality of the clustering results.

5.0 Wave Cluster

Wave Cluster is a multi resolution clustering algorithm. It is used to find clusters in very large spatial databases [5].

Given a set of spatial objects O_i, $1 \leq i \leq N$, the goal of the algorithm is to detect clusters. It first summarizes the data by imposing a multi dimensional grid structure on to the data space. The main idea is to transform the original feature by applying wavelet transform and then find the dense regions in the new space. A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub bands.

The first step of the wavelet cluster algorithm is to quantize the feature space. In the second step, discrete wavelet transform is applied on the quantized feature space and hence new units are generated. Wave cluster connects the components in 2 set of units and they are considered as cluster. Corresponding to each resolution γ of wavelet transform there would be set of clusters c_r, where usually at the coarser resolutions number of cluster is less? In the next step wave cluster labels the units in the feature space that are included in the cluster.

5.1 Advantages

1)Wavelet transformation can automatically result in the removal of outliers 2) Multi resolution properly of wavelet transformation can help in the detection of clusters at varying levels of accuracy based 3) Wavelet based clustering is very fast with a computational complexity of o(n) where n is the number of objects in the database 4) Discovers clusters with arbitrary shapes 5) It is insensitive to the order of input 6)It can handle data up to 20 dimensions 7) If can handle any large spatial database efficiently.

6.0 O-Cluster: (Orthogonal partitioning CLUSTERing)
This clustering method combines a novel partitioning active sampling technique with an axis parallel strategy to identify continuous areas of high density in the input space. O-cluster is a method that builds upon the contracting projection concept introduced by optigrid [6].

O-cluster makes two major contributions i) It uses statistical test to validate the quality of a cutting plane. This statistical test identifies good splitting points along data projections. ii) It can operate on a small buffer containing a random sample from the original data set. Partitions that do not have ambiguities are “frozen” and the data points associated with them are removed from the active buffer.

O-cluster operates recursively. It evaluates possible splitting points for all projections in a partition, selects the “best” one, and splits the data into new partitions.

The algorithm proceeds by searching for good cutting planes inside the newly created partitions. O-cluster creates a hierarchical tree structure that translates the input space into rectangular regions. The main processing stages are (1) Load data buffer (2) compute histograms for active partitions (3) Find “best” splitting points for active partitions (4) Flag ambiguous and “frozen” partitions (5) Split active partitions (6) Reload buffer.

O-cluster is a non parametric algorithm. O-cluster functions optimally for large data sets with many records & high dimensionality.

7.0 Axis Shifted Grid Clustering Algorithm (ASGC)

ASGC is a clustering technique which combines density and grid based methods to group objects with axis shifted partitioning strategy. The clustering quality of most of the grid based algorithms is influenced by the size of the predefined cells and the densities of the cells. This method uses two grid structures to reduce the impact of border of cells. The second grid structure is formed by shifting the coordinate axis by half a cell width in each dimension.

The method adapted by ASGC involves 6 steps i) The entire data space is divided into non overlapping cells thus forming first grid structure. ii) Significant cells are identified (If the density of the cells are more than predefined threshold). iii) All nearest significant cells are grouped together to form clusters. iv) Original coordinate origin is shifted by distance d in each dimension of the data space to get a new grid structure v) New clusters are generated using steps 2 and steps 3 vi) The clusters generated from both of the grid structures can be used to revise the clusters generated from the other grid structure.

ASGC algorithm has the advantage of low time complexity. It is a non parametric algorithm. This algorithm pre processes the data space and reduces the dimension of the data space.

8.0 Adaptive Mesh Refinement (AMR)

Adaptive Mesh Refinement is a type of multi resolution algorithm that achieves high resolution in localized regions [8]. Instead of using a single resolution mesh grid, AMR clustering algorithm first adaptively creates different

resolution grids based on the regional density. Secondly, the algorithm considers each leaf as the centre of an individual cluster and recursively assigns the membership for the data objects located in the parent nodes until the root node is reached. AMR Clustering algorithm can detect nested clusters at different levels of resolutions.

AMR is a technique that starts with a coarse uniform grid covering the entire computational volume and automatically refines certain regions by adding finer sub grids. New child grids are created from the connected parent grid cells whose attributes, density for instance, exceed given thresholds. Refinement is performed on each grid separately and recursively until all regions are captured with the desired accuracy.

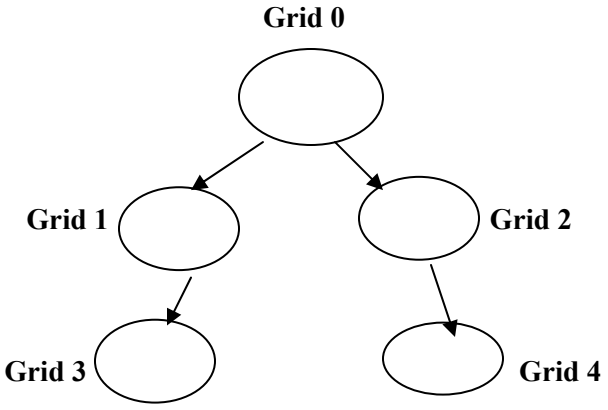

Figure -1 AMR Tree

Figure-1 shows an example of AMR tree in which each tree node uses a different resolution mesh. The root grid with the coarsest granularity covers the entire domain, which contains two sub grids, grid 1 and grid 2. Grid 2 at level 1 also contains two sub grids that are discovered using a finer mesh. The deeper the node is located in the tree, the finer the mesh is used.

AMR clustering algorithm connects the grid based and density based approaches through AMR techniques and hence preserves the advantages of both algorithms.

Comparison of Grid Based Algorithms

Algorithm Name	Type of Data	Complexity	Input Parameters
CLIQUE	**Numerical Data**	**$O(C^k+mk)$** **k – Highest Dimensionality** **m- number of input points** **C-number of clusters**	**ξ Number of Intervals** **τ – Density Threshold**
STING	**Spatial Data**	**O(K)** **K – Number of Cells at bottom layer**	**Number of objects in a cell**
MAFIA	**Numerical Data**	**$O(c^{k'}+N/p^B k'\gamma+ \alpha Spk')$** **B – Number of records that fit in memory buffer** **γ - I/O access time for a block of B records** **α – Constant for communication** **N – Total number of records** **S – Size of messages exchanger among processors** **P – Number of Processors** **k' – Number of Dimensions**	**-**
WAVE CLUSTER	**Spatial Data**	**O(n)** **n – number of objects**	**Wavelets, the number of grid cells for each dimension, the number of application of wavelet transform**
O CLUSTER	**Categorical and Numerical Data**	**O(N x d)** **N – number of objects** **d – Number of dimensions**	**Sensitivity Parameter**
ASGC	**Spatial Data**	**$O(m^d)+O(n)$** **n – number of data points** **d – sub set of attributes** **m – number of intervals**	**-**

9.0 Conclusion

In this research paper we have analyzed various grid based clustering algorithms and compared their complexity and input parameters. The recent algorithm such as Axis Shifted Grid Clustering Algorithm and Adaptive mesh refinement greatly reduces the impact of size of the grids and also improves the cluster quality.

10.0 References

[1] Aggarwal, C.C., Procopiuc, C., Wolf, J.L., Yu, P.S. and Jong S.P. (1999): fast Algorithms for projected Clustering. In Proc. of 1999 ACM SIGMOD International Conference on Management of Data, 61-72.

[2] Rakesh Agrawal, Johannes Gehrke, Dimirios Gunopulos, Prabhakar Raghavan: Automatic Subspace Clustering of High Dimensional Data. Data Mining and knowledge discovery, 11, 5-33, 2005. Springer Science + Business media, Inc. Manufactured in the Netherlands.

[3] Wei Wang, Jiong Yang, and Richard Muntz : STING : A Statistical Grid Appraoch to Spatial Data Mining : Department of Computer Science, University of California, Los Angels

[4] Sanjay Goil, Harsha Nagesh and Alok Choudhary : MAFIA: Efficient and Scalable Clustering for very large data sets : Technical Report No. CPDC – TR – 9906 – 010 ©1999 Center for Parallel and distributed Computing. June 1999

[5] Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang: WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases.

[6] Borinana L.Milenova, Marcos M.Campos: O-Cluster: Scalable Clustering of Large High Dimensional Data sets. Oracle Data Mining Technologies.

[7] Chung I-Chang, Nancy P Lin, Nien Yi Jan: An Axis Shifted Clustering Algorithm, Tamkang journal of Science and Engineering, Vol.12, No.2, pp.183-192 (2009)

[8] Wei-Keng Liao, Ying Liu, Alok Choudhary: A Grid-based Clustering Algorithm using Adaptive Mesh Refinement. Appears in the 7th Workshop on Mining Scientific and Engineering Data Sets 2004