D DURGAN BHAVANI, et al , International Journal of Computers Electrical and Advanced Communication Engineering [IJCEACE]™
Volume 1 , Issue 2, PP: 231 - 236 , NOV' 2012.

HIERARCHICAL CLUSTERING APPROACH USING QUAD TREE DECOMPOSITION FOR SPATIAL DATA

D DURGAN BHAVANI 1*, Dr. V VIJAY KUMAR 2*

1. Research Scholar, Dept of CSE, Acharya Nargarjuna University..
2. Dean & Prof, Dept of CSE, GIET, Rajahmundry.

Abstract: Quad tree is used to describe a class of hierarchical data structure whose common property is that they are based on the principle of recursive decomposition. Spatial data decomposition and clustering is a descriptive task that seeks to identify homogeneous groups of objects on the values of their attributes. This paper specifically focuses on clustering on multidimensional mixed category data. With the gridded representation of a multidimensional dataset, the present paper progresses on spatial clustering techniques. The Quad tree, a prevalent spatial data structure for representing two dimensional data based on regional homogeneity, is used for decomposition.

Keywords: Mixed multi dimensional data, clustering, and Quad tree decomposition.

I.INTRODUTION

Knowledge extraction in databases is the non trivial process of identifying valid patterns from the data. This process predicts the trends and behavior of large accumulated data to make a positive driven decision. There are many steps that are involved in assessing the decision like searching for patterns, knowledge evaluation, refinement and minimization of redundancy. The main steps that were involved in data mining or knowledge extraction is given in figure 1

The process starts with the understanding the application domain and identifying the goal of the process from the user's perspective, the selected data may contain the noises or some missing values so in the pre processing stage these sought of noises are cleaned to make the data more consistent .In the next step the data is transformed in more consolidated form appropriate for mining operations like normalization and aggregation. Data redundancy is one major criterion which makes the data to be stored repeatedly at different intervals, minimizing it enables us to choose the useful features.

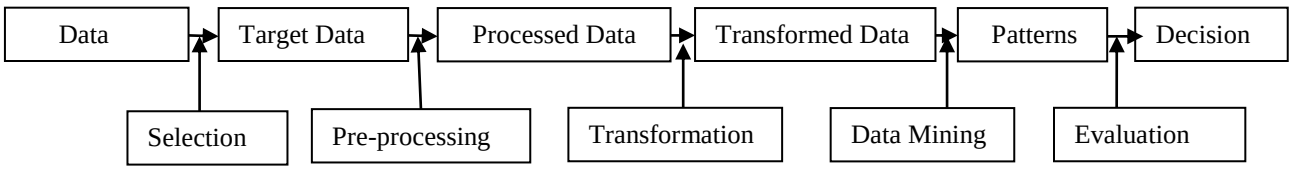

Figure 1: Block diagram of knowledge based data mining

In the next step of data mining, does the exploratory analysis on the data and builds some model based on known techniques of statistics and pattern recognition algorithms. This step includes applying the data analysis and discovery algorithms on the preprocessed subsamples and the transformed data. It is the application of specific algorithms for extracting patterns from the data and allows fitting a model for it. Data mining approaches can be categorized into descriptive based data mining and predictive based data mining, the formed describes the data in concise and the latter constructs one or a set of models to perform on it. Finally a decision will be taken based the evaluation of the extracted patterns.

II.BACKGROUND

A lot of research was done on data mining clustering approaches as clustering is primary subject in this paper , a few concepts which are related to the present context are presented.

Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes [1]. For the spatial data, clustering permits a generalization of the spatial component like explicit location and extension of spatial object which define implicit relations of spatial neighborhood. Spatial clustering techniques are categorized into partitioned, hierarchical and density based algorithms. In this paper more focus on hierarchical based algorithms are discussed.

Hierarchical based algorithms

A data set is said to be hierarchical cluster if there exists 2 samples c1 & c2 which are in the same cluster at some level 'k' and remain clustered together at all higher levels. The hierarchy is represented as a tree called dendrogram with individual elements at one end and a single cluster containing every element at the other. Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms called hierarchical agglomerative clustering, treat each object as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster. Top-down or divisive clustering proceeds by splitting clusters recursively until individual objects s are reached.

Agglomerative algorithms

(i) CURE (clustering using representatives) identifies clusters having non-spherical shapes and wide variances in size [2]. CURE is a bottom-up hierarchical clustering algorithm, but instead of using a centriod-based approach. In fact, CURE begins by choosing a constant number, 'c' of well scattered points from a cluster. These points are used to identify the shape and size of the cluster.

(ii) ROCK (Robust clustering using links) implements a new concept of links to measure the similarity/proximity between a pair of data points [3]. A pair of data points are considered neighbors if their similarity exceeds a certain threshold. The number of links between a pair of points is then the common neighbors for the points. Points

belonging to a single cluster will have a large number of common neighbors.

III. Quad Tree Decomposition

Spatial data mining or knowledge discovery in spatial databases differs from regular data mining in parallel with the differences between non-spatial data and spatial data. The attributes of a spatial object stored in a database may be affected by the attributes of the spatial neighbors of that object. In addition, spatial location and implicit information about the location of an object may be exactly the information that can be extracted through spatial data mining [4].

The term quad tree is used to describe a class of hierarchical data structure whose common property is that they are based on the principle of recursive decomposition of space [5].

Spatial data can be differentiated on the following bases: (1) the type of data that they are used to represent, (2) the principle guiding the decomposition process, and (3) the resolution

The decomposition may be partitioning the data into equal parts on each level, or unequal parts as may be governed by the requirement. The resolution of the decomposition (i.e., the number of times that the decomposition process is applied) may be fixed before hand or it may be governed by properties of the input data. Figure 2 shows the quad tree decomposition .

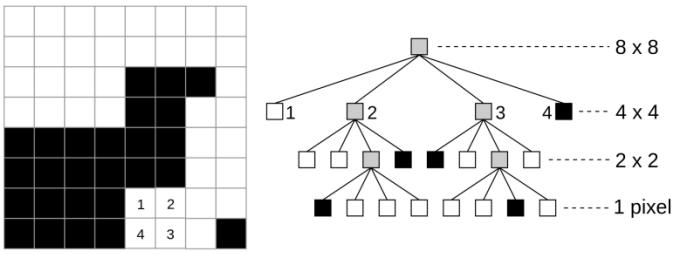

Figure 2: Quad tree decomposition

Process of decomposition

Quad tree decomposition is an analysis technique that involves subdividing an image into blocks that are more homogeneous. For the ease of computation, the image is square in shape with its length *n* is considered to be of the power of 2. The top-down approach will consist of successive non overlapping subdivisions of the original image blocks with dimensions 2x2 by a factor of four. If a block is homogeneous, given some criterion it is not subdivided. If it is non-homogeneous, it is subdivided into non-overlapping four sub-blocks. Quad tree encoding will segregate the image into equal 4 quadrants sub-blocks. All these sub-division of the non-homogeneous blocks will continue until the smallest block reaches a minimum pre-established block size. The pixel values of the image frame is normalized to a range of [0 1].

Decomposition based on Homogeneity

A region is said to be homogeneous if, if all the pixels in the region are within a specific dynamic range. For example, in a black and white image, a region is homogeneous if it entirely white or entirely black or minute enough to be identified as black or white.

The mean and the variance are two statistical measures which can be effectively used to find the homogeneity of the region. Each Quad tree region is analyzed for its mean and variance in pixel values. Entropy is also a statistical measure of randomness that can be used to characterize the texture of the input image. Lower entropy implies higher homogeneity and vice versa.

$$\mu = \frac{1}{MxN}\sum_{0}^{M-1}\sum_{0}^{N-1} f(x,y) \quad (1)$$

$$\sigma = \frac{1}{MxN}\sqrt{(f(x,y)-\mu)^2} \quad (2)$$

IV RESULT ANALYSIS

Different types of spatial datasets are used to evaluate the performance of the quad tree decomposition. The spatial data is first decomposed at 'N' levels with quad tree decomposition an then the clustering analysis is applied

(a) Iris dataset:

Iris dataset is a classic data set with 50 samples of iris flowers measured in four attributes, sepal length and width, petal length and width. Cluster analysis on this data set only contains two clusters with rather obvious separation. One of the clusters contains *Iris setosa*, while the other cluster contains both *Iris virginica* and *Iris versicolor* and is not separable without the species information [6].

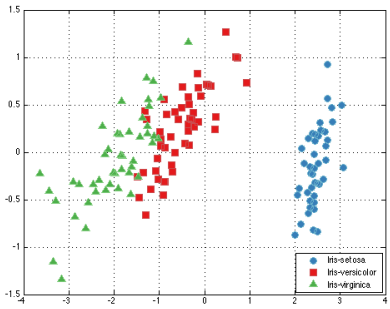

Figure 3: output of the clustered iris data

(b) Yeast dataset:

Yeast dataset contained 1484 instance of gene expression data defined by 8 predictive localized site of protein. The remapping of cluster indices after the fourth phase of the framework grouped the data in to 2 clusters; one main cluster containing 99% of the data and another cluster with feeble number of values.

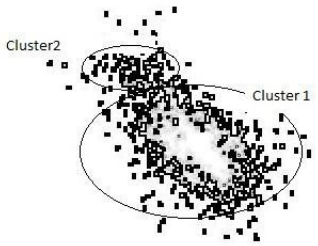

Figure 4: output of the clustered yeast data

Quad tree based cluster in spatial domain

For the evaluated a natural image is considered which is decomposed using quad tree and clustered for object segmentation

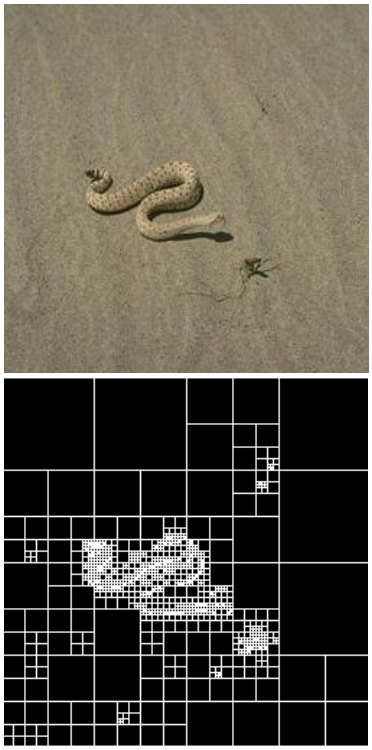

Figure 5: (a) original Image (b) Clustered Image

V. CONCLUSION

A quad tree decomposed based image clustering is shown for image segmentation using the data mining concepts. Experimental results show that the present decomposition approach not only applies for normal data but also implements a segmentation application for the spatial data. This work can be further extended with more precise hierarchical algorithms for clustering .

REFERENCES

[1] Ester, M., Frommelt, A., Kriegel, H.-P., and Sander, J. Algorithms for characterization and trend detection in spatial databases. In: Int. Conf. on Knowledge Discovery and Data Mining; 1998; New York City, NY. p. 44-50.

[2] Sudipto Guha , Rajeev Rastogi , Kyuseok Shim. CURE: an efficient clustering algorithm for large databases. In: Proceedings of the 1998 ACM SIGMOD international conference on Management of data; 1998; Seattle. p. 73-84.

[3] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim. ROCK: A Robust Clustering Algorithm for Categorical Attributes. In: Proceedings of the 15th international Conference on Data Engineering; 1999. p. 512-521.

[4] Usama Fayyad, Gregory Piatetsky-shapiro,Padhraic Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. 1996. p. 82-88.

[5] Samet H. The Quadtree and Related Hierarchical Data Structures. ACM Computing Surveys. 1984;16(2):187-260.

[6] http://en.wikipedia.org/wiki/Iris_flower_data_set

[7] He Q. A Review of Clustering Algorithms as Applied in IR [Internet]. 1999.

[8] Ming-Yi Shih, Jar-Wen jheng, Lien-Fu Lai. A two step method for Clustering Mixed Categorical and Numeric Data. Tamkang Journal of Science and Engineering. 2010;13(1):11-19.

[9] Samet H. The Quadtree and Related Hierarchical Data Structures. ACM Computing Surveys. 1984;16(2):187-26

[10] Qixiang Ye ; Wen Gao ; Wei Zeng. Color image segmentation using density-based clustering. In: ICASSP '03; 2003. p. 345-348.