A Fuzzy Grid-Clustering Algorithm

[1]K. Yogeswara Rao, [2]Ch. Sita Kameswari, [3]D. Siva Phanindra
[1,2]Dept. of MCA, Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, Andhra pradesh
[3]Dept. of CSE, St.Ann's College of Engineering, Chirala, Andhra Pradesh

Abstract
A grid clustering algorithm normally partitions the data space into a finite number of cells to form a grid structure and then performs all clustering operations to group similar spatial objects into classes on this grid structure. In this paper, we explore fuzzy grid clustering algorithm where an object need not belong to a single class. Instead, we define a quantitative value which represents the degree up to which the object is a member of the class. This algorithm is used to cluster efficiently and simultaneously to reduce the size and borders of cells. This new method is a combination of standard fuzzy logic and grid clustering.

Key Words
Data Mining, Grid-Based Clustering, Significant Cell, Grid Structure, Coordinate Axis, Fuzzy Logic

I. Introduction

Clustering analysis which is to group the data points into clusters is an important task of data mining in recent days. Unlike classification which analyzes the labeled data, clustering analysis deals with data points without consulting a known label previously. In general, data points are grouped only based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity, and thus, clusters of data points are formed so that data points within a cluster are highly similar to each other, but are very dissimilar to the data points in other clusters. Up to now, many clustering algorithms have been proposed, and generally, the so called grid-based algorithms are the most computationally efficient ones. The main procedure of the grid-based clustering algorithm is to partition the data space into a finite number of cells to form a grid structure, and next, find out the significant cells whose densities exceed a predefined threshold, and group nearby significant cells into clusters finally. Clearly, the grid-based algorithm performs all clustering operations on the generated grid structure; therefore, its time complexity is only dependant on the number of cells in each dimension of the data space. That is, if the number of the cells in each dimension can be controlled as a small value, then the time complexity of the grid-based algorithm will be low. Some famous algorithms of grid-based clustering are STING [1], STING+ [2], WaveCluster [3], and CLIQUE [4].

As mentioned above, the grid-based clustering algorithm is an efficient algorithm, but its effect is seriously influenced by the size of the grids (or the value of the predefined threshold). And the weakness of continuity in cells is in the border. So how to select the borders of cells is another important issue. If the cell is small, then it needs many cells to be connected into one cluster. And there will also be more connection of cells. In the connection of cells, the number of data points in cell is the major factor to connect or disconnect the cells. So, the more cells the more effects. And in the same data space, there are more cells, there will be smaller size. To cluster data points efficiently and to reduce the influences of the size of the cells at the same time, a new grid-based clustering algorithm, the Fuzzy Grid-Clustering algorithm (FGC) is proposed here.

The main idea of FGC is to reduce the impact of border of cells by using two grid structures. FGC shifts the original grid structure in each dimension of the data space after the clusters generated from the original grid structure have been obtained. The shifted grid structure is then used to find out the new significant cells. Next, the nearby significant cells are grouped as well to form some new clusters. Finally, these new generated clusters are used to revise the originally generated clusters.

The rest of the paper is organized as follows: In section II, An introduction to the theory of Fuzzy logic is given. In section III, the proposed clustering algorithm, the Fuzzy Grid-Clustering algorithm, will be presented. In section IV, some experiments and discussions will be displayed. The conclusions will be given in section V.

II. Fuzzy Logic

In general, if we wish to say whether an object belongs to a certain class, we would represent it by either 0' or '1'.'0' would indicate that the object does not belong to the class. '1' would indicate that the object is completely inside the class. In Fuzzy Logic this is not the case. Every object would have to be processed with a membership function and assign a degree of membership (DOM). This DOM indicates how much is the object related to particular class. So instead of '0' and '1', an object could now have DOM as 0.5,0.8 and so on.

A. Membership Function

A membership function determines the DOM of an object. Triangular, Trapezoidal, Gaussian etc are some of the membership functions that can be used. A Triangular Membership function is shown below.

Tra (x, a, b, c)=

$$\begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a \le x \le b \\ \frac{c-x}{c-b}, & b \le x \le c \\ 0, & c \le x \end{cases}$$

'x' is the density of the cell
'a' is the minimum density of the cells in the cluster.
'b' is the median density of the cells in the cluster.
'c' is the maximum density of the cells in the cluster

III. Fuzzy Clustering Algorithm

This is infact an extension of the ASGC algorithm.After the first grid structure is built, the algorithm shifts the coordinate axis by half a cell width in each dimension and has the new grid structure, then combines the two sets of clusters into the final result.

Step 1: Generate the first grid structure.

By dividing into k equal parts in each dimension, the n dimensional data space is partitioned into k^n non-overlapping cells to be the first grid structure.

Step 2: Identify significant cells.

Next, the density of each cell is calculated to find out the significant

cells whose densities exceed a predefined threshold.

Step 3: Generate the set of clusters

We use membership functions to determine the degree to which a significant cell is a member of a cluster for this we use in membership function on the density of each cell. This gives the degree of membership (DOM) for the cell in that cluster. The process to perform this is as follows :

First, as in the original fuzzy grid clustering algorithm, near by significant cells which are connected to each other are grouped into clusters.

For each cell in the cluster, we apply any of the membership function on it and find the DOM of the cell in that cluster.

This cell is again considered for its degree of membership in other cluster. In short, though we define clusters with each containing a subset of cells, we represent every cell in every cluster with its DOM.

Step 4: Transform the grid structure

The original co-ordinate origin is next shifted by distance d in each dimension d in each dimension of the data space, so that the coordinate of each point becomes d less in each dimension.

Step 5: Generate set of new clusters

The step 2 and step 3 are used repeatedly to generate the set of new clusters by using the transformed grid structure. The set of new clusters generated here is denoted as S_2.

Step 6: Revise original clusters

The clusters generated from the second grid structure can be used to revise the originally obtained clusters. And the first grid structure can also be used to revise the second obtained clusters. The procedure of Revision of the original clusters is shown in Fig. 1.

1) Find each overlapped cluster C_{2j} for $C_{1i} \in S_1$

2) Generate the rule $C_{1i} \rightarrow C_{2j}$,
where $C_{1i} \cap C_{2j} \neq \phi$, $C_{2j} \in S_2$

3) Find each overlapped cluster C_{1i} for $C_{2j} \in S_2$

4) Generate the rule $C_{2j} \rightarrow C_{1i}$,
where $C_{2j} \cap C_{1i} \neq \phi$.

5) $R_o = \{C_{1i} \rightarrow C_{2j} \mid C_{1i} \cap C_{2j} \neq \phi\} \cup \{C_{2j} \rightarrow C_{1i} \mid C_{2j} \cap C_{1i} \neq \phi\}$

6) $C_{1i} \in S_1$ are revised by using the cluster revised function CM().

Fig.1 : The sketch of Revision of the original clusters.

Step 6 A: Find each overlapped cluster C_{2j} for C_{1i} $\in S_1$, and generate the rule $C_{1i} \rightarrow C_{2j}$, where $C_{1i} \in S1$, $C_{2j} \in S_2$. The rule $C_{1i} \rightarrow C_{2j}$ means that cluster C_{1i} overlaps cluster C_{2j}. Similarity, find each overlapped cluster C_{1i} for $C_{2j} \in S_2$, and also generate the rule $C_{2j} \rightarrow C_{1i}$

Step 6 B: The set of all rules generated in the step 6a is denoted as R_0. Next each cluster $C_{1i} \in S_1$ is revised by using the cluster revised function CM(). The cluster modified function CM() is shown in fig. 1.

Step 7: Generate the clustering result.

After all clusters of S_1 have been revised, S_2 is the rest of the original set of S_2 after revision. The final set of clusters is $S_1 = S_1 \rho S_2$. The result will be the same as S_2 revised by S_1.

IV. Experiment Results

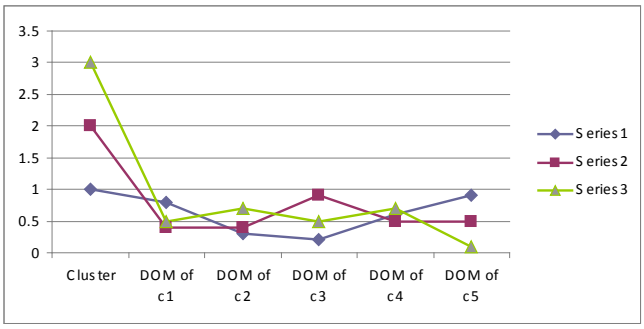
To demonstrate the algorithm we here take five cells c1,c2,c3,c4,c5 and apply the above the algorithm. The result is tabulated below.

Table 1: Experimental data features

Cluster	DOM of c1	DOM of c2	DOM of c3	DOM of c4	DOM of c5
1	0.8	0.3	0.2	0.6	0.9
2	0.4	0.4	0.9	0.5	0.5
3	0.5	0.7	0.5	0.7	0.1

The above table shows how Fuzzy grid clustering algorithm assigns a DOM to each of the cells.

The graph shows for the above experimental data features.



V. Conclusion

In this paper, the new Fuzzy grid clustering algorithm, has been introduced. This algorithm proves to be effective in associating a cell to every cluster with certain DEGREE OF MEMBERSHIP. There are some helpful technique can be used in FGC algorithm. One is to use the non-parametric algorithm to find the first one fitted size and density threshold to get the natural clustering results. And another is to use the technique of factor analysis to preprocess the data space and reduce the dimension of data space.

References

[1]. Wang W., Yang J., Richard, R., Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," In Proc. of 23rd Int. Conf. on VLDB, pp. 186Ę195 (1997).

[2]. Wang W., Yang J., Richard, R., Muntz, "STING+: An Approach to Active Spatial Data Mining," In Proc. of 15th Int. Conf. on Data Engineering, pp. 116Ę125 (1999).

[3]. Sheikholeslami, G., Chatterjee, S., Zhang, A., "WaveCluster: A Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases," In VLDB Journal: Very Large Data Bases, pp. 289Ę304 (2000).

[4]. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., "Automatic Sub-Space Clustering of High Dimensional Data for Data Mining Applica-tions," In Proc. of ACM SIGMOD Int. Conf. MOD, pp. 94Ę105 (1998).

K. Yogeswara Rao MCA, M.Tech, is working as an Assistant Professor in the Department of Computer Applications at Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, India. He has 13 years of teaching experience. His area of interest in Data Mining, Data base Management System. He has guided more than 50 Post Graduate Students. He has published 2 books, 1 paper in international Conference and has attended 10 National Workshops / FDP / Seminars etc.

Ch. Sita Kameswari MCA, M.Tech, MBA, (Ph.D), is working as an Associate Professor in the Department of Computer Applications at Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, India. She has 13 years of teaching experience. Her area of interest is Artificial Intelligence, Computer Graphics, and Machine Learning. She has guided more than 50 Post Graduate Students. She has published 1 paper in international Journal & 3 papers in proceedings of International conference including 1 IEEE conference and has attended 12 National Workshops / FDP / Seminars etc.

D. Siva Phanindra, B. Tech is working as an Assistant Professor in the Dept. of Computer Science Engineering, St.Ann's college of engineering, Chirala, India. He has 13 years of teaching experience. His area of interest in Image Processing, Data Mining, Data base Management System.