Accurate Grid-based Clustering Algorithm with Diagonal Grid Searching and Merging

Feng Liu¹, Chengcheng Ye¹ and Erzhou Zhu¹, a)

¹School of Computer Science and Technology, Anhui University, Hefei 230601, China
E-mail: a) ezzhu@ahu.edu.cn

Abstract. Due to the advent of big data, data mining technology has attracted more and more attentions. As an important data analysis method, grid clustering algorithm is fast but with relatively lower accuracy. This paper presents an improved clustering algorithm combined with grid and density parameters. The algorithm first divides the data space into the valid meshes and invalid meshes through grid parameters. Secondly, from the starting point located at the first point of the diagonal of the grids, the algorithm takes the direction of “horizontal right, vertical down” to merge the valid meshes. Furthermore, by the boundary grid processing, the invalid grids are searched and merged when the adjacent left, above, and diagonal-direction grids are all the valid ones. By doing this, the accuracy of clustering is improved. The experimental results have shown that the proposed algorithm is accuracy and relatively faster when compared with some popularly used algorithms.

1. Introduction
As an important means in data mining, clustering algorithm is widely used in many fields such as data analysis, pattern recognition, image processing and so on. By clustering, sample points in each cluster have high similarity while different cluster sample points with low similarity.

By now, many kinds of clustering algorithms are emerged. By calculating the distance between the points and the cluster, the partition-based clustering algorithm, likes K-Means [1], is suitable for finding convex clusters. Unavailable of the number of clusters in advance, hierarchical clustering algorithm, likes BIRCH [2], is clustered by computing the distance between each cluster pair. This algorithm is suitable for aggregating multiple local clusters, but cannot deal with large-scale data properly.

By iteratively merging neighborhood high-density regions, Density-based clustering algorithms can identify arbitrary shapes data sets. But the time cost of this method is huge. DBSCAN [3] is a spatial clustering algorithm based on high density connectivity. It can find any shape of the cluster in the noisy data. However, it is very sensitive to the parameters ε (radius) and MinPts (neighborhood density threshold).

Grid-based clustering algorithms divide the data space into finite units and then cluster them. CLIQUE [4] proposed by Agrawal is a representative algorithm based on density and grid. It can automatically identify the subspace of the high dimensional data space. However, the need for finding the clustering of all the lower subspaces reduces the time and space performance of the algorithm. TSING [5] is another grid clustering algorithm with high performance but the accuracy is relatively low.

In order to elevate the accuracy of the grid clustering algorithm, this paper proposes DSM, an improved grid-based clustering algorithm with diagonal grid searching and merging. At the beginning,
the data space is meshed by the grid parameter. If the points in the grid are greater than or equal to the threshold parameter, it is treated as a valid grid; otherwise it is considered an invalid grid. The algorithm takes the first point of the diagonal is selected as the starting grids. It also adopts horizontal-right and vertical-downward direction along the diagonal to search the valid grids. Furthermore, the valid grids that are superjacent or on the left of the valid grids located at horizontal and vertical directions of the diagonal are merged into the same groups. This search strategy can effectively reduce the time cost. During the search process, the boundary grids are judged. If it is an invalid grid, then expand searching the left, above, and diagonal meshes of this invalid grid. If they are all valid grids, then directly merging them for clustering accuracy elevates.

2. The implementation of dsm
The algorithm proposed in this paper aims at improving the accuracy and efficiency of the grid-based clustering algorithm. This method is based on dividing the grid into space and then dividing them into "valid grid" or "invalid grid" according to the grid containing at least the predetermined threshold parameter. By taking the first point of the diagonal of grids as starting grids, the effective grids are searched along the horizontal and vertical directions of the diagonal. During the searching process, if the grids adjacent to the searched valid grids are already clustered, these grids can be treated as the same group. Otherwise, they are put into a new group. Finally, if the left adjacent grid and the upper adjacent grid contain multiple groups, all groups will be merged into the same group.

In order to illustrate the detailed steps of the algorithmic method, it is needed to specify some parameters in advance: the DataSet represents the data set, GridSize represents the size of the grids, and MinPts represents the least data points in the grid. In this algorithm, NeighborGridCheck() and GridGroupMerge() are two important modules. NeighborGridCheck() will check whether the adjacent left and top of the grid are valid and have already been clustered. If so, they can be merged into the same group. If the adjacent left, top and diagonal of the grid are invalid grid, it may be treated as a new group. In addition, GridGroupMerge() is responsible for merging the current grid and the group where the neighbors are located, and merging all the groups into the same group. As Fig.1 shows, the proposed algorithm has the following steps:

**Step1:** Initialize all parameters.

**Step2:** Divide the grid space and distinguish between the valid grids and invalid grids according to the grid parameter and threshold parameter to. As shown in Fig.2.
FIGURE 2. Dividing grid space into valid and invalid grids.

**Step 3:** Use the diagonal grids as the starting grids to search valid grid in horizontal and vertical directions. If the current grid on the left or top of the searched grid located at the diagonal is the valid one, it is merged into its adjacent left or upper group. Otherwise, all the valid grids are merged into the same group.

As shown in Fig.3, the diagonal grid 1 horizontal search to the effective grid 2. If 1 and 3 are valid grid, then 1, 2 and 3 are in the same group. Otherwise, set to a new group. Grid 1 vertical down search to grid 4. If 1 and 5 are valid grids, then the grids 1, 4 and 5 are in the same group. So 1, 2, 3, 4 and 5 are in the same group.

FIGURE 3. The process of searching and merging grids.

**Step 4:** In the search and merge process, if the left, top and diagonal adjacent grids are valid, the grid is determined as the boundary grid and merged it into the group of the left, top and diagonal grid.

As shown in Fig.4, invalid grid 2 extended search grids 1, 2 and 3. If the grids 1, 2 and 3 are valid grids, then merge together. Fig.5 shows the clustering results.

FIGURE 4. Merge the invalid grid.

FIGURE 5. Clustering result.
3. Experiment and analysis
In this work, the algorithm is implemented in Java program and executed in the desktop computer with 4GB RAM, Intel Core i3 3.7 GHz on Windows 10 operating system. The clustering results are compared with the DBSCAN, CLIQUE, and TSING methods. According to Fig.6, the data sets 1 to 4 have different distribution patterns with 575,000 data points and 75,000 noise points. Data set 1, 2, 3 and 4 has 5, 10, 3 and 4 clusters respectively [6].

The data sets 1 to 4 were clustered 10 times using the proposed algorithm and the conventional method, and the average results are listed in Table 1. Fig.6 shows the clustering results. It can be seen from Table 1 that the comparison between the proposed algorithm and other methods shows that the clustering accuracy is improved by extending the boundary void grid. Meanwhile, the only search three adjacent grids per grid improves the time efficiency. DSM consumes more time but exhibits more clustering accuracy when compared with TSING.

![FIGURE 6. The original datasets and clustering results of datasets 1-4.](image)

**TABLE1.** Experimental results; item1 indicates run-time cost(s); item 2 indicates clustering accuracy rate(%).

| Algorithm | Item | DataSet-1 | DataSet-2 | DataSet-3 | DataSet-4 |
|-----------|------|-----------|-----------|-----------|-----------|
| DBSCAN    | 1    | 27495     | 27545     | 27483     | 27479     |
|           | 2    | 99.99%    | 99.98%    | 99.99%    | 99.99%    |
| TSING     | 1    | 0.052     | 0.057     | 0.055     | 0.053     |
|           | 2    | 98.62%    | 99.58%    | 99.45%    | 99.49%    |
| CLIQUE    | 1    | 0.563     | 0.556     | 0.570     | 0.573     |
|           | 2    | 98.85%    | 99.72%    | 99.60%    | 99.64%    |
| DSM       | 1    | 0.432     | 0.437     | 0.454     | 0.448     |
|           | 2    | 98.93%    | 99.79%    | 99.70%    | 99.80%    |

4. Conclusion
In this paper, a grid-based algorithm is proposed to improve the accuracy of grid clustering algorithm. It takes the diagonal grids as starting grids to horizontal and vertical search and merge valid grids. For each grid, only two adjacent grids are searched. As a result, the proposed data clustering algorithm can reduce the cost of clustering time and improve the efficiency. In addition, the processing of the boundary value is added to improve the clustering accuracy.

References
[1] J. B. McQueen, “Some Methods of Classification and Analysis of Multivariate Observations,” the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281-297.
[2] T. Zhang and R. Ramakrishnan, “BIRCH: An efficient Data Clustering Method for Very Large Databases,” ACM SIGMOD International Conference on Management of Data, 1996, pp. 103-114.
[3] M. Ester, H. P. Kriegel and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226-231.

[4] R. Agrawal, J. E. Gehrke, D. Gunopulos and P. Raghavan, “Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications,” ACM SIGMOD International Conference on Management of Data, 1998, pp. 94-105.

[5] C. F. Tsai and Y. C. Hu, “Enhancement of efficiency by thrifty search of interlocking neighbor grids approach for grid-based data clustering,” International Conference on Machine Learning & Cybernetics, 2013, pp. 1279-1284.

[6] C. F. Tsai and S. C. Huang, “An effective and efficient grid-based data clustering algorithm using intuitive neighbor relationship for data mining,” International Conference on Machine Learning & Cybernetics, 2015, pp. 478-483.