An Axis-Shifted Grid-Clustering Algorithm

Chung-I Chang[1]*, Nancy P. Lin[2] and Nien-Yi Jan[2,3]

[1]*Department of Information Management, St. Mary's Medicine, Nursing and Management College*
[2]*Department of Computer Science and Information Engineering, Tamkang University,*
Tamsui, Taiwan 251, R.O.C
[3]*Business & Marketing Strategy Research Department, Telecommunication Lab., Chunghwa Telecom Co.,*
Ltd Taiwan, R.O.C

Abstract

These spatial clustering methods can be classified into four categories: partitioning method, hierarchical method, density-based method and grid-based method. The grid-based clustering algorithm, which 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 obtained grid structure, is an efficient clustering algorithm. To cluster efficiently and simultaneously, to reduce the influences of the size and borders of the cells, a new grid-based clustering algorithm, an Axis-Shifted Grid-Clustering algorithm (ASGC), is proposed in this paper. This new clustering method combines a novel density-grid based clustering with axis-shifted partitioning strategy to identify areas of high density in the input data space. The main idea is to shift the original grid structure in each dimension of the data space after the clusters generated from this original structure have been obtained. The shifted grid structure can be considered as a dynamic adjustment of the size of the original cells and reduce the weakness of borders of cells. And thus, the clusters generated from this shifted grid structure can be used to revise the originally obtained clusters. The experimental results verify that, indeed, the effect of this new algorithm is less influenced by the size of cells than other grid-based ones and requires at most a single scan through the data.

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

1. Introduction

Clustering analysis which is to group the data points into clusters is an important task of data mining recently. 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 [1–8], and generally, the 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 the

*Corresponding author. E-mail: taftdc@smc.edu.tw grid-based clustering are STING [1], STING+ [2], WaveCluster [3], and CLIQUE [4].

As the above mentioned, 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 Axis-Shifted Grid-Clustering algorithm (ASGC) is proposed here.

The main idea of ASGC is to reduce the impact of border of cells by using two grid structures. ASGC 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 2, some famous grid-based clustering algorithms will be introduced. In section 3, the proposed clustering algorithm, the Axis-Shifted Grid-Clustering algorithm, will be presented. In section 4, some experiments and discussions will be displayed. The conclusions will be given in section 5.

2. Grid-Based Clustering Algorithm

In this section, six popular grid-based clustering algorithms, STING [1], STING+ [2], CLIQUE [4], GDILC [5], NSGC [6], GCHL [7], and ADCC [8] will be introduced.

STING and STING+ exploit the clustering properties of index structures. They employ a hierarchical grid structure and use longitude and latitude to divide the data space into rectangular cells. They select a layer to begin with at the beginning.

For each cell of this layer, to label the cell as relevant if its confidence interval of probability is higher than the threshold. We go down the hierarchy structure by one level and go back to check those cells is relevant or not until the bottom level. Return those regions that meet the requirement of the query. And finally, to retrieve those data fall into the relevant cells.

CLIQUE [4] is a density and grid-based approach for high dimensional data sets that provides automatic sub-space clustering of high dimensional data. It consists of the following steps: First, to uses a bottom-up algorithm that exploits the monotonicity of the clustering criterion with respect to dimensionality to find dense units in different subspaces. Second, it use a depth-first search algorithm to find all clusters that dense units in the same connected component of the graph are in the same cluster. Finally, it will generate a minimal description of each cluster.

GDILC [5] is a novel clustering algorithm. The central idea of GDILC is that the density-isoline figure depicts the distribution of data samples very well. It uses a grid-based method to calculate the density of each data sample, and find relatively dense regions, which are just clusters. GDILC is capable of eliminating outliers and finding clusters of various shapes. It is an unsupervised clustering algorithm because it requires no human interaction.

NSGC [6] is a new density- and grid-based type clustering algorithm using the concept of shifting grid is proposed. The proposed algorithm is a non-parametric type, which does not require users inputting parameters. It divides each dimension of the data space into certain intervals to form a grid structure in the data space. Based on the concept of sliding window, shifting of the whole grid structure is introduced to obtain a more descriptive density profile. It clusters data in a way of cell rather than in points.

GCHL [7] combines a novel density-grid based clustering with axis-parallel partitioning strategy to identify areas of high density in the input data space. The algorithm works as well in the feature space of any data set. The method operates on a limited memory buffer and requires at most a single scan through the data. GCHL has the high quality of the obtained clustering solutions, capability of discovering concave/deeper and convex/higher regions, their robustness to outlier and noise, and excellent scalability.

ADCC [8] is the first grid-based algorithm to shift the original grid structure in each dimension of the data space after the clusters generated from this original structure have been obtained. But in the step to identify the set of clusters in both generations, it checks all cells not only the significant cells.

3. Axis-Shifted Grid-Clustering Algorithm

In fact, the effects of most grid-based algorithms are seriously influenced by the size of the predefined grids and the threshold of the significant cells. And the continuity of border in cells is the weakness of grid-based clustering methods. To reduce the influences of the border of the predefined grids and increase the selection of size and density threshold of the significant cells and also improve the result of clustering, we propose the ASGC algorithm in this paper.

After the first grid structure is built, the ASGC 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. The procedure of ASGC is shown in Figure 1.

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 who se densities exceed a predefined threshold.

Step 3: Generate the set of clusters.
Then the nearby significant cells which are connected to each other are grouped into clusters. The set of clusters is denoted as S_1.

Step 4: Transform the grid structure.
The original coordinate origin is next shifted by distance d in each dimension of the data space, so that the coordinate of each point becomes d less in each dimension.

Step 5: Generate the set of new clusters.
The step 2 and step 3 are used again 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 Figure 2.

Step 6a: Find each overlapped cluster C_{2j} for $C_{1i} \in S_1$, and generate the rule $C_{1i} \rightarrow C_{2j}$, where $C_{1i} \cap C_{2j}$

Generate the first grid structure.
Identify the significant cells.
Generate the first set of clusters.
Transform the grid structure.
Generate the second set of new clusters.
Revise the original clusters.
Generate the clustering result.

Figure 1. The ASGC algorithm.

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().

Figure 2. the sketch of Revision of the original clusters

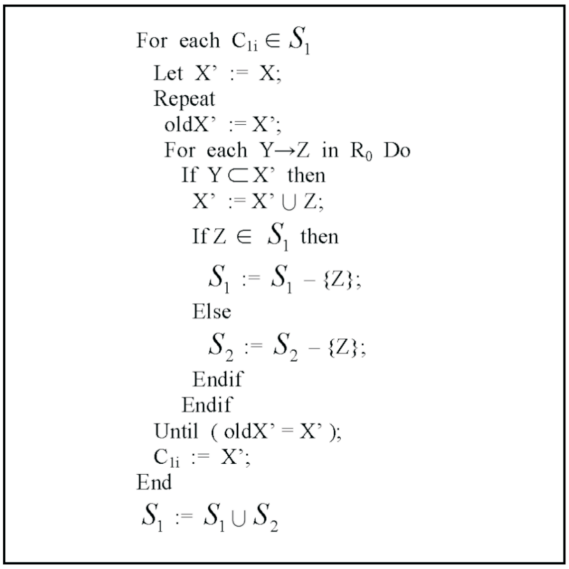
$\neq \phi$, $C_{2j} \in S_2$. The rule $C_{1i} \rightarrow C_{2j}$ means that cluster C_{1i} overlaps cluster C_{2j}. Similarity, find each overlapped cluster C_{1i} for $C_{2j} \in S_2$, and also generate the rule $C_{2j} \rightarrow C_{1i}$, where $C_{2j} \cap C_{1i} \neq \phi$.

Step 6b: The set of all the rules generated in step 6a is denoted as R_o. Next, each cluster $C_{1i} \in S_1$ is revised by using the cluster revised function *C*M(). The cluster modified function *C*M() is shown in Figure 3.

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 $S_1 = S_1 \cup S_2$. The result will be the same as S_2 revised by S_1.

In this place, the two-dimensional example is easy to figure out and understand, as shown in Figure 4, with 799 points is easy to be divided into four natural clusters. The example goes through the ASGC algorithm.

At first, the two-dimensional data space in this example is partitioned into 20^2 non-overlapping cells to be the grid structure. Next, the density of each cell is calculated to find out the significant cells whose densities exceed the predefined threshold, here the threshold is 5. Then the nearby significant cells which are connected to each other are grouped into 11 clusters. The first set of clusters is denoted as $S_1 = \{C_{11}, C_{12}, \ldots, C_{1\,11}\}$, shown in Figure 5.



```
For each $C_{1i} \in S_1$
  Let X' := X;
  Repeat
   oldX' := X';
   For each $Y \rightarrow Z$ in $R_0$ Do
     If $Y \subset X'$ then
      $X' := X' \cup Z$;
      If $Z \in S_1$ then
        $S_1 := S_1 - \{Z\}$;
      Else
        $S_2 := S_2 - \{Z\}$;
      Endif
     Endif
  Until ( oldX' = X' );
  $C_{1i}$ := X';
End
$S_1 := S_1 \cup S_2$
```

Figure 3. The CM algorithm.

The original coordinate origin is next shifted by half side length of the cell, so that the coordinate of each point becomes *d* (= half side length of the cell) less in each dimension. After shifting the grid structure, the data space is partitioned into 21^2 cells. Here, the cell density of new grid structure is also calculated. It's easy to find out the significant cells whose densities exceed the predefined threshold, 5. And the nearby significant cells which are connected to each other are grouped into 14 clusters. The second set of the clusters is denoted as $S_2 = \{C_{21}, C_{22}, \ldots, C_{2\,14}\}$, as shown in Figure 6.

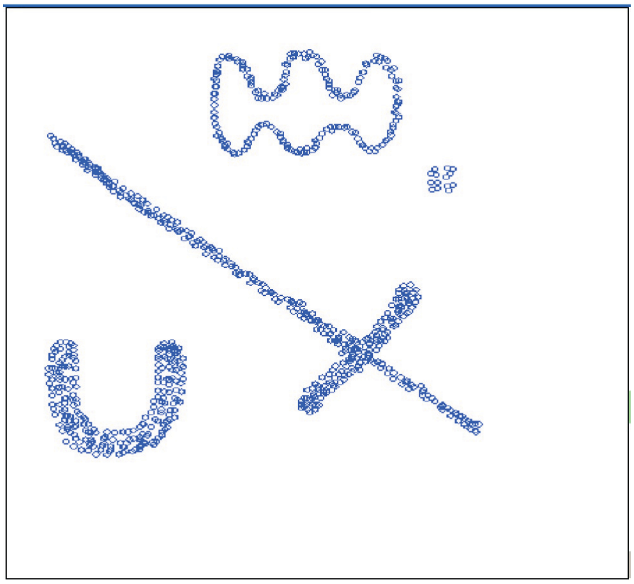


Figure 4. Original data in the two-dimensional data space

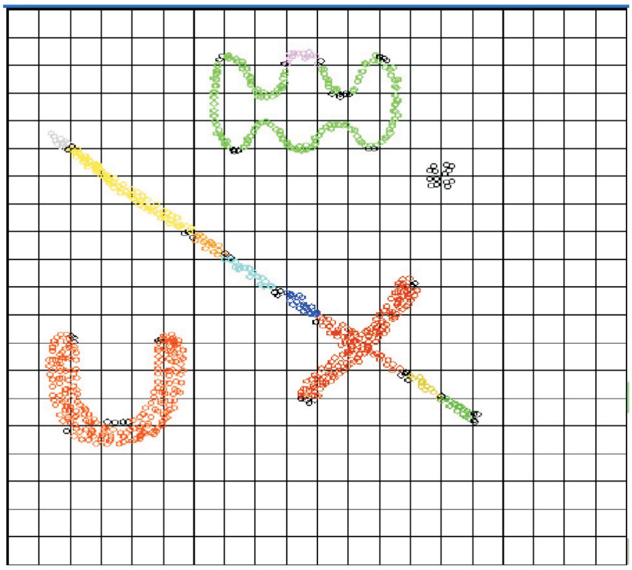


Figure 5. The first clustering S_1 (20^2 cells, 11 clusters).

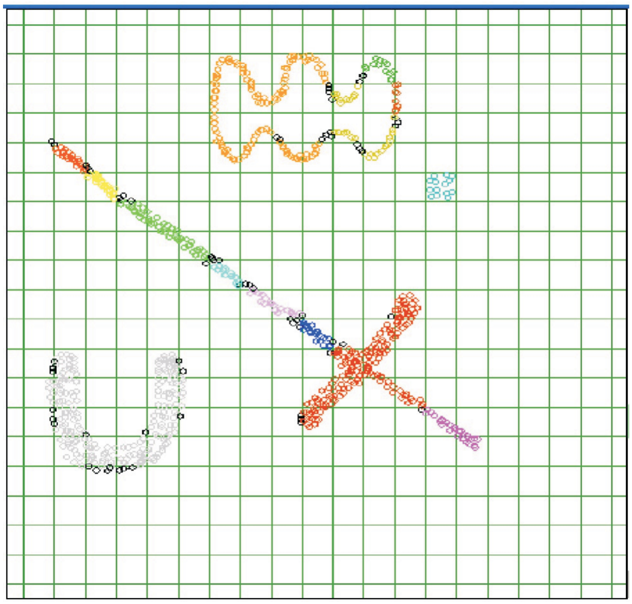


Figure 6. The second clustering S_2 (21^2 cells, 14 clusters).

The clusters generated from the shifted grid structure are used to revise the originally obtained clusters. R0 is composed of rules $C_{1i} \rightarrow C_{2j}$, shown in Table 1, and $C_{2j} \rightarrow C_{1i}$, shown in Table 2.

After all clusters of S_1 have been revised by using cluster modified function *C*M(), revised S_1 is shown in Table 3. And the final clustering result is shown in Figure 7.

4. Experiments

Here, we experiment with six different sets of data, as shown in Table 4. Owing to the feature of the two-dimensional examples, it is easy to figure out and understand the seven two-dimensional experiments, as shown in Figures 8 to 14. In each experiment, we randomly chose in the range from (1,16) to (5,55) for the total of 100 sets of different combination of the parameters (density threshold, number of dividing parts in each dimension). We tested not only by the ASGC algorithm we proposed, but also we used Hill-Climbing [9], K-means [10], and CLIQUE.

In Table 5, it is the correct rate comparison sheet of experiment 1 by using random 100 sets of parameters. The correct rate of the proposed algorithm (ASGC) is

Table 1. Rules $C_{1i} \rightarrow C_{2j}$ of R_0

S_1	Corresponding clusters in S_1	R_0 of S_1
1	2,3	$C_{11} \rightarrow C_{22}$, $C_{11} \rightarrow C_{23}$
2	1	$C_{12} \rightarrow C_{21}$
3	3,4	$C_{13} \rightarrow C_{23}$, $C_{13} \rightarrow C_{24}$
4	5,10,11,12	$C_{14} \rightarrow C_{25}$, $C_{14} \rightarrow C_{210}$ $C_{14} \rightarrow C_{211}$, $C_{14} \rightarrow C_{212}$
5	4,6	$C_{15} \rightarrow C_{24}$, $C_{15} \rightarrow C_{26}$
6	6,7	$C_{16} \rightarrow C_{26}$, $C_{16} \rightarrow C_{27}$
7	5	$C_{17} \rightarrow C_{25}$
8	7,8	$C_{18} \rightarrow C_{27}$, $C_{18} \rightarrow C_{28}$
9	8,9	$C_{19} \rightarrow C_{28}$, $C_{19} \rightarrow C_{29}$
10	9,14	$C_{110} \rightarrow C_{29}$, $C_{110} \rightarrow C_{214}$
11	14	$C_{111} \rightarrow C_{214}$

Table 2. Rules $C_{2j} \rightarrow C_{1i}$ of R_0

S_2	Corresponding clusters in S_2	R_0 of S_2
1	2	$C_{21} \rightarrow C_{12}$
2	1	$C_{22} \rightarrow C_{11}$
3	1,3	$C_{23} \rightarrow C_{11}$, $C_{23} \rightarrow C_{13}$
4	3,5	$C_{24} \rightarrow C_{13}$, $C_{24} \rightarrow C_{15}$
5	4,7	$C_{25} \rightarrow C_{14}$, $C_{25} \rightarrow C_{17}$
6	5,6	$C_{26} \rightarrow C_{15}$, $C_{26} \rightarrow C_{16}$
7	6,8	$C_{27} \rightarrow C_{16}$, $C_{27} \rightarrow C_{18}$
8	8,9	$C_{28} \rightarrow C_{18}$, $C_{28} \rightarrow C_{19}$
9	9,10	$C_{29} \rightarrow C_{19}$, $C_{29} \rightarrow C_{10}$
10	4	$C_{210} \rightarrow C_{14}$
11	4	$C_{211} \rightarrow C_{14}$
12	4	$C_{212} \rightarrow C_{14}$
13	X	{}
14	10,11	$C_{214} \rightarrow C_{110}$, $C_{214} \rightarrow C_{111}$

Table 3. The set of final clusters

New C_{1i}	Corresponding original C_{1i} and C_{2j}
C_{11}	C_{11}, C_{22}, C_{23}, C_{13}, C_{24}, C_{15}, C_{26}, C_{16}, C_{27}, C_{18}, C_{28}, C_{19}, C_{29}, C_{110}, C_{214}, C_{111}
C_{12}	C_{12}, C_{21}
C_{13}	-
C_{14}	C_{14}, C_{25}, C_{210}, C_{211}, C_{212}, C_{17}
C_{15}	-
C_{16}	-
C_{17}	-
C_{18}	-
C_{19}	-
C_{110}	-
C_{111}	-
C^*_{213}	C_{213}

C^*_{213}: in C_{213}, the data points are belong to a significant cell, which are partitioned into four non-significant cells in C_{1i}.

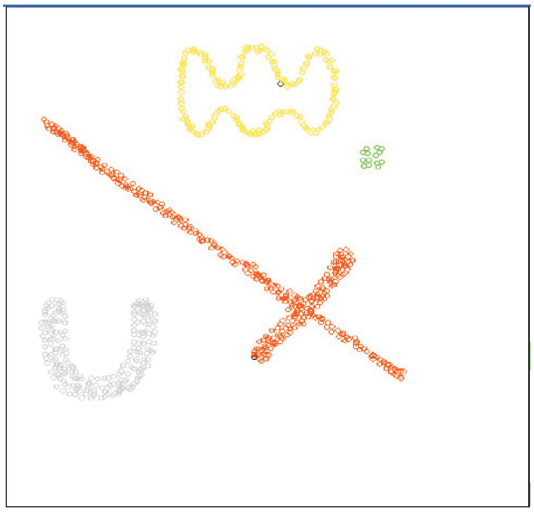


Figure 7. The final clustering result of ASGC.

Table 4. Experimental data features

Data	Number of Data	Natural clustering number
Exp 1	600	4
Exp 2	1100	4
Exp 3	1100	5
Exp 4	1150	4
Exp 5	900	3
Exp 6	1000	2
Exp 7	785	3

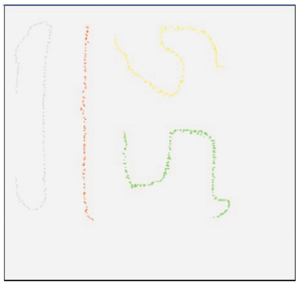


Figure 8. Experiment 1.

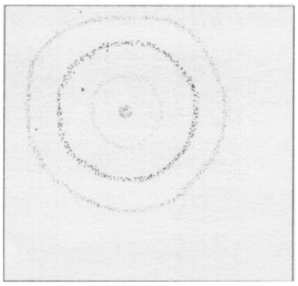


Figure 9. Experiment 2.

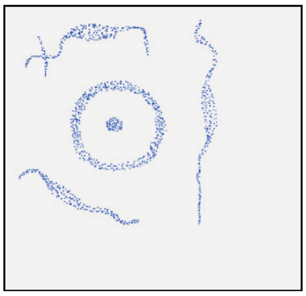


Figure 10. Experiment 3.

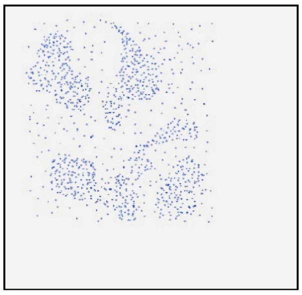


Figure 11. Experiment 4.

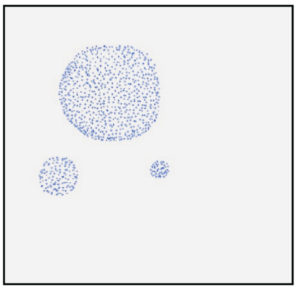


Figure 12. Experiment 5.

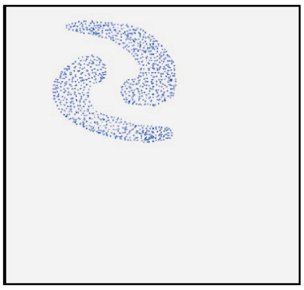


Figure 13. Experiment 6.

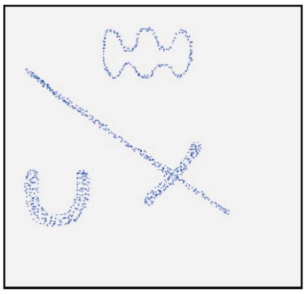


Figure 14. Experiment 7.

Table 5. The correct rate comparison sheet of experiment 1

CLIQUE ASGC	Correct clustering	Incorrect clustering	Correct Rate of ASGC
Correct clustering	**2%**	**45%**	**47%**
Incorrect clustering	**0%**	**53%**	
Correct Rate of CLIQUE	**2%**		

47% which is higher than CLIQUE whose correct rate is only 2%. Here, the correct result of CLIQUE is part of the clustering result of ASGC in experiment 1. And it does not find the wrong experimental result that using the ASGC but is correct when using CLIQUE.

From Table 6 to Table 11, it is possible to find the correct experimental result that using in the proposed algorithm but is wrong when using in CLIQUE. Because the correct rate of the ASGC algorithm is always higher than CLIQUE, the experiment by using the proposed algorithm is able to advance the correct rate than using other grid-based algorithms. In other words, the experimental results verify that the effect of the proposed algorithm is less influenced by the size of the cells than other grid-based ones.

Figure 15 shows the correct rates of the proposed al-

Table 6. The correct rate comparison sheet of experiment 2

CLIQUE ASGC	Correct clustering	Incorrect clustering	Correct Rate of ASGC
Correct clustering	**12%**	**55%**	**67%**
Incorrect clustering	**1%**	**32%**	
Correct Rate of CLIQUE	**13%**		

Table 7. The correct rate comparison sheet of experiment 3

CLIQUE ASGC	Correct clustering	Incorrect clustering	Correct Rate of ASGC
Correct clustering	**22%**	**42%**	**64%**
Incorrect clustering	**1%**	**35%**	
Correct Rate of CLIQUE	**23%**		

Table 8. The correct rate comparison sheet of experiment 4

CLIQUE ASGC	Correct clustering	Incorrect clustering	Correct Rate of ASGC
Correct clustering	**16%**	**44%**	**60%**
Incorrect clustering	**1%**	**39%**	
Correct Rate of CLIQUE	**1%**		

Table 9. The correct rate comparison sheet of experiment 5

CLIQUE ASGC	Correct clustering	Incorrect clustering	Correct Rate of ASGC
Correct clustering	**72%**	**9%**	**81%**
Incorrect clustering	**4%**	**15%**	
Correct Rate of CLIQUE	**76%**		

Table 10. The correct rate comparison sheet of experiment 6

CLIQUE ASGC	Correct clustering	Incorrect clustering	Correct Rate of ASGC
Correct clustering	**42%**	**31%**	**73%**
Incorrect clustering	**0%**	**27%**	
Correct Rate of CLIQUE	**42%**		

Table 11. The correct rate comparison sheet of experiment 7

CLIQUE ASGC	Correct clustering	Incorrect clustering	Correct Rate of ASGC
Correct clustering	**7%**	**55%**	**62%**
Incorrect clustering	**0%**	**38%**	
Correct Rate of CLIQUE	**7%**		

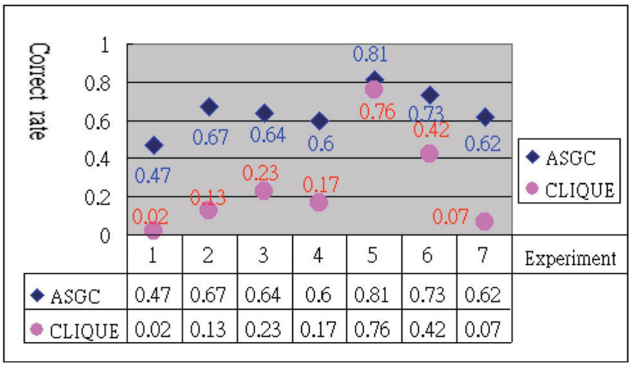


Figure 15. Correct rates of CLIQUE and ASGC.

gorithm (ASGC) and CLIQUE. The correct rates of the proposed algorithm are all higher than CLIQUE. In the experiments, the correct rates comparison is by using random 100 sets of parameters (density threshold, number of dividing parts in each dimension) from (1,16) to (5,55).

We tested not only by the ASGC and CLIQUE algorithms, but also we used Hill-Climbing and K-mean.

The comparisons of the correct clusters and running time are shown in Table 12, the ASGC algorithm inherits the advantage with the low time complexity. The clustering results of Hill_Climbing, K-mean and CLIQUE almost are not as good as ASGC. The Hill_Climbing has the advantage with the low time complexity, but the results of clustering are bad. The shortcomings of the K-mean algorithm are its tendency to favor spherical clusters with similar size and number of data, and the fact that the knowledge on the number of clusters, *k*, is required in advance.

4. Discussion

In the experiments, we tested not only by the ASGC algorithm we proposed, but also we used Hill-Climbing, K-mean, and CLIQUE.

Table 12. average executing time and number of clusters among four algorithms

Original Data		Hill-climbing		K-mean		CLIQUE		ASGC	
NO.	Natural clusters	Cluster	time (ms)	Cluster	time (ms)	Cluster	time (ms)	Cluster	time (ms)
1	4	24	40	4	1180	4	60	4	170
2	4	42	55	4	611	4	80	4	230
3	5	21	68	5	922	5	100	5	380
4	4	15	90	4	1680	4	102	4	260
5	3	11	44	3	850	3	62	3	220
6	2	11	62	2	1640	2	90	2	350
7	3	17	35	3	570	3	50	3	170

The Hill-Climbing method [9] is a density-based clustering algorithm. It has the tendency to group the cells from the valley to the peak. The valleys will be the boundary of clusters. In Table 12, the hill-climbing method has the least time, but the number of clusters is more than the natural number of clusters. The Figures 16 and 17 are the clustering results of Hill-Climbing and ASGC.

The K-mean method [10] is one of the famous clustering algorithms. It has the tendency to group the points with the least distance to the center of cluster they belong to. The number of clusters is proper and predefined, but the clustering result of data with uneven shapes sometimes is improper. In Figure 18, each point belongs to the cluster that its center is the nearest one with the point in the figure. The Figures 18 and 19 are the clustering results of K-mean and ASGC.

Table 13 and Table 14 show the number of acceptable density threshold and divided parts in each dimension of the CLIQUE and ASGC algorithms. Both the acceptable range of parameters in ASGC is wider than the acceptable range of parameters in CLIQUE.

Table 15 shows the acceptable pairs (m, d), m is number of divided parts and d is the number of density threshold. In the experiments, the ASGC is always better than CLIQUE.

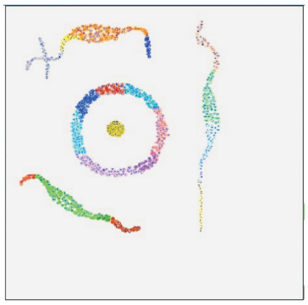


Figure 16. Hill-Climbing.

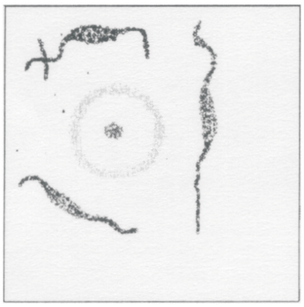


Figure 17. ASGC.

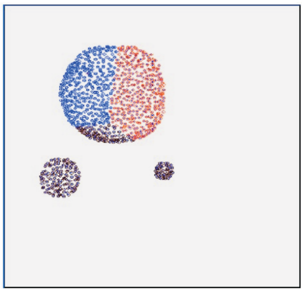


Figure 18. K-mean.

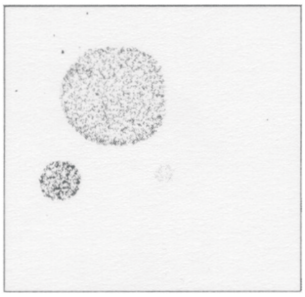


Figure 19. ASGC.

Table 13. Comparison of CLIQUE and ASGC (d: density threshold)

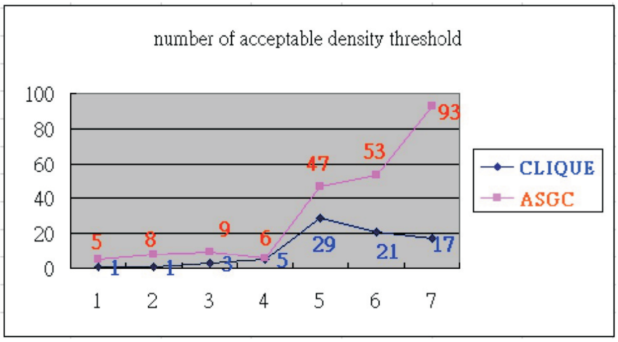


Table 14. Comparison of CLIQUE and ASGC (m: number of divided parts)

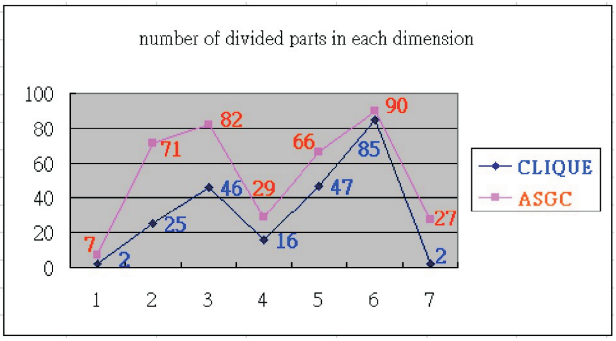
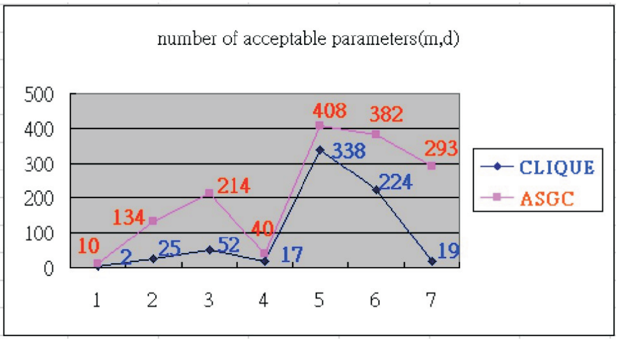


Table 15. Comparison of CLIQUE and ASGC (m, d)



Depending on the results of Hill-Climbing and CLIQUE, the incorrect result happened by the borders in cells and bad parameters. And the step 4 in the ASGC algorithm is to transform the grid structure. Owing to the two revise of clusters, revised by original or shifted grid structure, the coordinate origin of the data space is chosen randomly and removes the impact of border in cells. The shifted grid structure also can be considered as a dynamic adjustment of the size of the original cells and reduce the weakness of borders of cells. The size of cell is reduced to be $1/2^d$ of the original cell but with strong continuity.

In the ASGC algorithm, for each data sample α, only those samples that are in the same cell of α are considered. The density of such cell is calculated. When the number of data samples is n and each dimension, total d dimensions, is divided into m intervals, there will be m^d cells. The time of checking the density of all cells is $k0 * [m^d + (m+1)^d]$. If $p (= 3^d - 1)$ is the number of nearby cells of one cell, the time of checking the cell is significant or not is $k1 * p * [m^d + (m + 1)^d]$ at most. So the time of Revision of the original clusters in ASGC is $k2 * [m^d + (m + 1)^d]$ at most. In the end, the time of checking the cluster's number of all data is $k3 * n$. So the total time complexity is $O(m^d) + O(n)$.

5. Conclusion

In this paper, the new grid-based clustering algorithm, the Axis-Shifted Grid-Clustering algorithm, has reduced the weakness of border in cells and increases the obvious wider ranges of size of the cell and threshold of density. And the experimental results verify that the effect of ASGC algorithm is less influenced by the size of the cells than other grid-based ones. Owing to the two revise of clusters, revised by original or shifted grid structure, the final results of clustering are the same, the coordinate origin of the data space is chosen randomly and removes the impact of border in cells. At the same time, the ASGC algorithm still inherits the advantage with the low time complexity and requires at most one single scan through the data.

There are some helpful technique can be used in ASGC 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. and Richard, R., Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," *In Proc. of 23rd Int. Conf. on VLDB*, pp. 186–195 (1997).

[2] Wang W., Yang J. and Richard, R., Muntz, "STING+: An Approach to Active Spatial Data Mining," *In Proc. of 15th Int. Conf. on Data Engineering,* pp. 116–125 (1999).

[3] Sheikholeslami, G., Chatterjee, S. and Zhang, A., "WaveCluster: A Wavelet-Based Clustering Approach for Spatial Data in Very Large Databases," *In VLDB Journal: Very Large Data Bases,* pp. 289–304 (2000).

[4] Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P., "Automatic Sub-Space Clustering of High Dimensional Data for Data Mining Applications," *In Proc. of ACM SIGMOD Int. Conf. MOD,* pp. 94–105 (1998).

[5] Zhao, Y. C. and Song, J., "GDILC: A Grid-Based Density-Isoline Clustering Algorithm," *In Proc. Internat. Conf. on Info-net,* Vol. 3, pp. 140–145 (2001).

[6] Ma, W. M., Eden, Chow and Tommy, W. S., "A New Shifting Grid Clustering Algorithm," *Pattern Recognition,* Vol. 37, pp. 503–514 (2004).

[7] Pilevar, A. H. and Sukumar, M., "GCHL: A Grid-Clustering Algorithm for High-Dimensional Very Large Spatial Data Bases," *Pattern Recognition Letters,* Vol. 26, pp. 999–1010 (2005).

[8] Lin, Nancy P., Chang, C.-I., Chueh, H.-E., Chen, H.-J. and Hao, W.-H., "An Adaptable Deflect and Conquer Clustering Algorithm," *In Proceedings of the 6th*

WSEAS International Conference on Applied Computer Science, pp. 155–159 (2007).

[9] Russell, Stuart J. and Norvig, Peter, Artificial Intelligence: A Modern Approach (2nd ed.), Upper Saddle River, NJ: Prentice Hall, pp. 111–114 (2003).

[10] MacQieen, J., "Some Methods for Classification and Analysis of Multivariate Observation," *Proc. 5th Berkeley Symp. Math. Statist, Prob.,* Vol. 1, pp. 281–297 (1967).

Manuscript Received: Nov. 21, 2007
Accepted: Jul. 24, 2008