# An Adaptive Crossover-Imaged Clustering Algorithm

NANCY P. LIN[1], CHUNG-I CHANG[1], HAO-EN CHUEH[1],
HUNG-JEN CHEN[1,2], WEI-HUA HAO[1]
[1]Department of Computer Science and Information Engineering
Tamkang University
151 Ying-chuan Road Tamsui, Taipei County
TAIWAN, R.O.C
[2]Department of Industrial Engineering and Management
St. John's University,
499, Sec. 4, Tam-King Road, Tamsui, Taipei,
TAIWAN, R.O.C.

*Abstract*: - The grid-based clustering algorithm is an efficient clustering algorithm, but its effect is seriously influenced by the size of the predefined grids and the threshold of the significant cells. The data space will be partitioned into a finite number of cells to form a grid structure and then performs all clustering operations on this obtained grid structure. To cluster efficiently and simultaneously, to reduce the influences of the size of the cells and inherits the advantage with the low time complexity, an Adaptive Crossover-Imaged Clustering Algorithm, called ACICA, is proposed in this paper. The main idea of ACICA algorithm is to deflect the original grid structure in each dimension of the data space after the image of significant cells generated from the original grid structure have been obtained. Because the deflected grid structure can be considered a dynamic adjustment of the size of original cells and the threshold of significant cells, the new image generated from this deflected grid structure will be used to revise the originally obtained significant cells. Hence, the new image of significant cells is projected on the original grid structure to be the crossover image. Finally the clusters will be generated from this crossover image. The experimental results verify that, indeed, the effect of ACICA algorithm is less influenced by the size of the cells than other grid-based algorithms. Finally, we will verify by experiment that the results of our proposed ACICA algorithm outperforms than others.

*Key-Words*: - Data Mining, Grid Structure, Crossover Image, Significant Cell, Deflected Grid

## 1 Introduction

Up to now, many clustering algorithms have been proposed [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14], 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 grid-based clustering are STING [11], WaveCluster [12], CLIQUE [13], and ADCC [14].

In general, grid-based clustering algorithm is the most computationally efficient algorithm, but the effect of grid-based clustering algorithm is seriously influenced by the size of the predefined grids and the threshold of the significant cells. To reduce the influences of the size of the predefined grids and the threshold of the significant cells, we propose a new grid-based clustering algorithm which is called Adaptive Crossover-Imaged Clustering (ACICA) algorithm in this paper.

The main idea of our proposed ACICA is to utilize some predefined grids and a predefined threshold to identify the image of significant cells in the first grid sructure. Then, the modified grids which are deflected to half size of the grid are used

to identify the image of significant cells in the second grid structure again. Next, the two images of significant cells are crossovered to generate the final clustering result.

The rest of the paper is organized as follows: In section 2, some popular grid-based clustering algorithms are mentioned again. In section 3, our proposed clustering algorithm, ACICA algorithm, is introduced. In section 4, an experiment and some discussions are displayed. Section 5 is the conclusion.

## 2 Grid-based Clustering Algorithm

Grid-based clustering algorithm is an efficient clustering algorithm, and three famous grid-based clustering algorithms are STING [11], CLIQUE [13], and ADCC [14].

STING (Statistical Information Grid-based algorithm) (Wang et al., 1997) is a grid-based clustering technique. It employs a hierarchical structure of grid cells and uses longitude and latitude to divide the spatial space into rectangular grid cells. Each cell at a high level is partitioned to form a number of cells at the next lower level. Statistical information regarding the attributes in each grid cell (such as the number of data, mean, maximum, minimum, and distribution values) is precomputed and stored. These statistical parameters are useful for query processing. At first, it selects a layer to begin with. Then, for each cell of this layer, to label the cell as relevant if its confidence interval of probability is higher than the threshold. Next, it goes down the hierarchy structure by one level and goes back to check those cells is relevant or not until the bottom level. Afterwards, return those regions that meet the requirement of the query. And finally, to retrieve those data fall into the relevant cells.

The CLIQUE (Clustering In QUEst) (Agrawal et al., 1998) clustering algorithm integrates density-based and grid-based clustering. For high dimensional data sets, it provides automatic sub-space clustering of high dimensional data. It consists of the following steps: First, to uses a bottom-up algorithm to find dense units in different subspaces. The CLIQUE based on the Apriori property that if a k-dimensional unit is dense, then so are its projections in (k-1)-dimensional space. Second, it uses 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.

The idea of ADCC (Adaptable Deflect and Conquer Clustering) (Lin et al., 2007) is to utilize the predefined grids and predefined threshold to identify the significant cells, by which nearby cells that are also significant can be merged to develop a cluster in the first place. Next, the modified grids which are deflected to half size of the grid are used to identify the clusters again. Finally, the new generated clusters and the initial clusters are merged to be the final clustering result.

In fact, to reduce the influences of the size of the predefined grids and the threshold of the significant cells, a new grid-based clustering algorithm proposed is called Adaptive Crossover-Imaged Clustering Algorithm (ACICA) in this paper.

## 3 The Adaptive Crossover-Imaged Clustering Algorithm

The Adaptive Crossover-Imaged Clustering (ACICA) algorithm further deflects the grid structure by half a cell width in each dimension, the same as ADCC algorithm, but ADCC generates the final clustering by using clustering procedure three times. To improve the chief defect, the crossover-imaged clustering is proposed. The ACICA generates the final clustering results only using one clustering and no more memory.

### 3.1 ADCC algorithm

After the grid structure is built, the ADCC deflects the cell margins by half a cell width in each dimension and have the new grid structure and then combine the two sets of clusters into the final result. The procedure of ADCC is shown in the following steps.

Step 1: Generate a 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 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.

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

Step 4: Deflect the grid structure.

The original grid structure is next deflected by distance $d$ in each dimension of the data space.

Step 5: Generate the set of new clusters.
The step 2 and step 3 are used again to generate the new set of clusters by using the deflected grid structure. The set of new clusters generated here is denoted as $S_2$.

Step 6: Revise original clusters.
The clusters generated from the deflected grid structure are used to revise the originally obtained clusters as the following steps.

Step 6a: Find each overlapped cluster $C_{2j}$ for $C_{1i} \in S_1$, and generate the rule $C_{1i} \rightarrow C_{2j}$, where $C_{1i} \bigcap C_{2j} \neq \phi, C_{2j} \in S_2$. 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 rules $C_{2j} \rightarrow C_{1i}$, where $C_{2j} \bigcap C_{1i} \neq \phi$.

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 $CR()$. The cluster modified function $CR()$ is shown in fig.1.

Step 7: Generate the clustering result.
After all clusters of $S_1$ have been revised, $S_1$ is the set of final clusters.

### 3.2 ACICA algorithm

After the two grid structures are built, the ACICA projects the second image of significant cells on the original grid structure to be the crossover image. Finally the clusters will be generated from this crossover image. The procedure of ACICA is shown in the following steps.

Step 1: Generate a 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 grid structure.

Step 2: Identify significant cells.
Next, the density of each cell is calculated to find out the image of significant cells whose densities exceed a predefined threshold.

Step 3: Deflect the grid structure.
The original grid structure is next deflected by distance $d$ in each dimension of the data space.

Step 4: Identify significant cells.
Next, the density of each cell is calculated to find out the new image of significant cells whose densities exceed a predefined threshold.

Step 5: Generate the crossover image
The second image of significant cells is generated and projected on the original grid structure to be the crossover image.

```
for each C1i ∈ S1
    Let X' := C1i;
    Repeat
      oldX' := X';
      For each Y→Z in R0 Do
          If Y ⊂ X' then
             X' := X' ∪ Z;
             If Z ∈ S1 then
                S1 := S1 – {Z};
          Endif
    Until (oldX' = X');
    C1i := X';
End
```

Fig.1 the CR algorithm

Step 6: Generate the clustering result.
Find each $OC^1_{X_1..Xn}$, the set of overlapped significant cells $C^2{}_{X1..Xn}$ for $C^1{}_{X1..Xn} \in G1$, grid structure 1, where $C^2{}_{X1..Xn}$ is the significant cell in the second n-dimensional grid structure. $OC^1_{X_1..Xn} = \{C^2{}_{X1..Xn} \mid C^2{}_{X1..Xn} \bigcap C^1{}_{x1..xn} \neq \phi\}$, where $C^2{}_{X1..Xn}$ is significant cells in G2, the second grid structure. By using the crossover image, the nearby significant cells in the original grid structure are clusrtered. The clusters are combined into one cluster depending on the cater-corner significant cells, which are connected each other by one corner. And the the set of cells, $OC^1_{X_1..Xn}$ overlapped the cluster in the original grid structure is combined into the same cluster.

### 3.3 Example

In this place, the two dimensional example, as shown in figure 2, with 1100 points is easy to be divided into two clusters. The example goes through the algorithm.

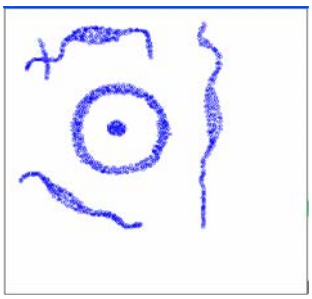

Fig. 2 figure of example

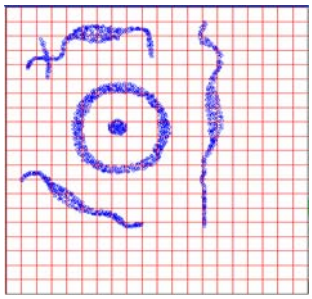

Fig.3 the grid structure

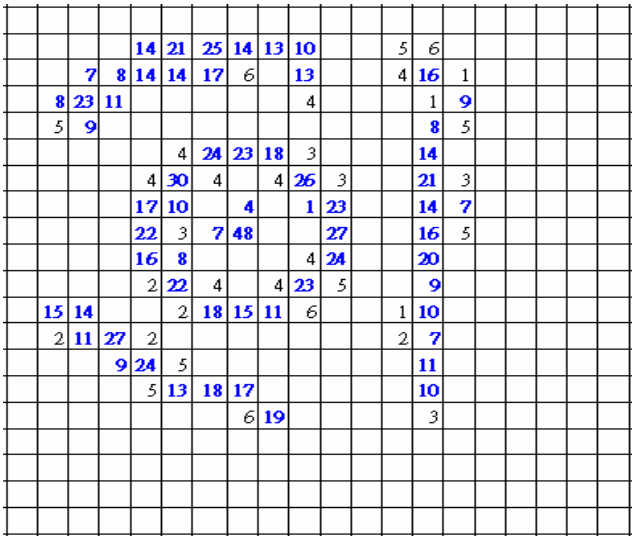

Fig.4 density of each cell

Step 1: Generate a grid structure.

By dividing into 20 equal parts in each dimension, the two dimensional data space in this example is partitioned into $20^2$ non-overlapping cells to be the grid structure, shown in fig.3.

Step 2: Identify significant cells.

Next, the density of each cell is calculated, shown in fig. 4, to find out the significant cells whose densities exceed a predefined threshold, here the threshold is 7. And the image of significant cells is shown in fig.5.

Step3: Deflect the grid structure.

The original grid structure is deflected by distance $d$ in each dimension of the data space. In this example, $d$ is equal to the half side length of the cell. By deflecting the grid structure, the new one is partitioned into $21^2$ cells, shown in fig. 6.

Step 4: Identify significant cells.

Here, the cell's density of new grid structure is shown in fig. 7. It's easy to find out the significant cells whose densities exceed a predefined threshold, 7.

Step 5: Generate the crossover image

The second image, shown in fig.8, of significant cells is projected on the original grid structure to be the crossover image shown in fig.9.

Step 6: Generate the clustering result.

The clustering result by using the crossover image is shown in fig.10.

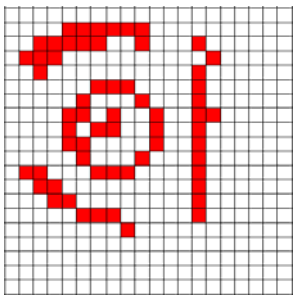

Fig.5 image of significant cells

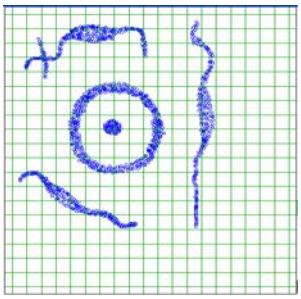

Fig.6 grid structure

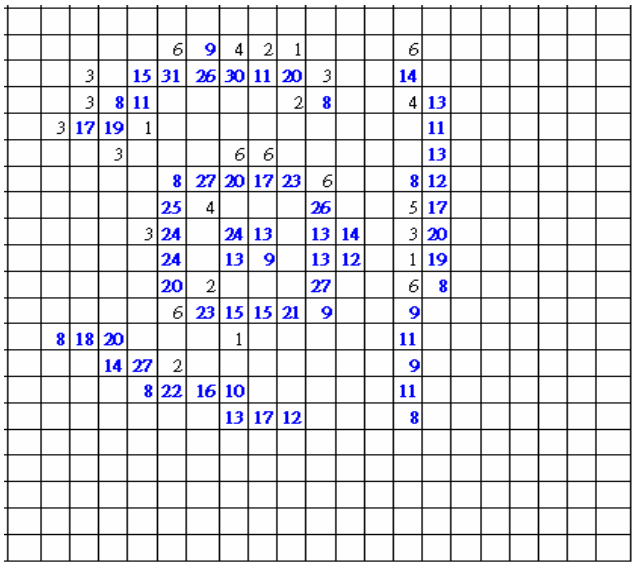

Fig.7 Density of each cell

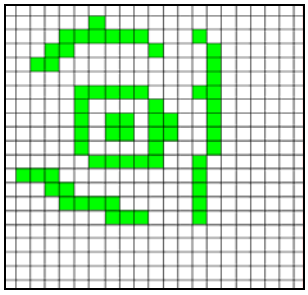

Fig.8 image of significant cells

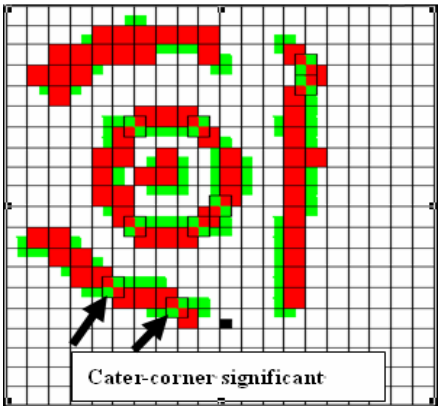

Fig.9 crossover image of significant cells

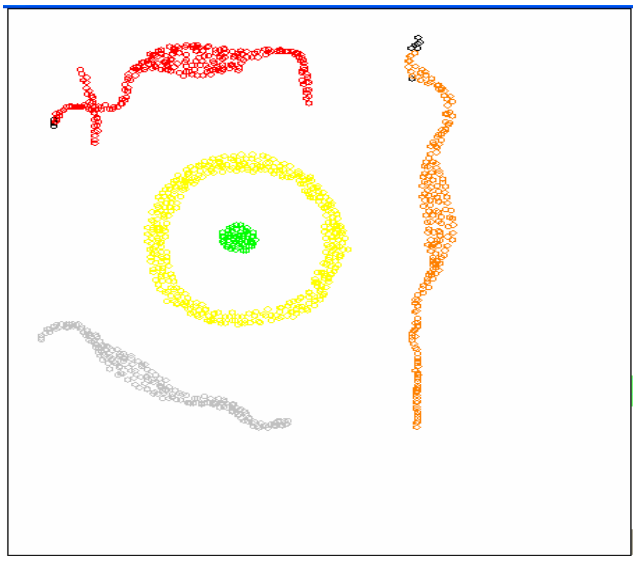
Fig.10 the clustering result

## 4. Experiments and Discussions

Here, we experiment with seven different data shown in fig.11 ~ fig.17. And the features are presented in Table 1.

| 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 |

Table 1 experimental data features

### 4.1 . Experiment

Figure.18 shows the correct rates of ACICA and SDG, where the correct clustering result of SDG is by using one of original or new grid structures in the experiment. The correct rates of ACICA are all higher than SDG. In the experiment, the correct


Fig.11 experiment 1

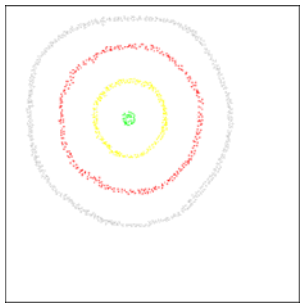
Fig.12 experiment 2

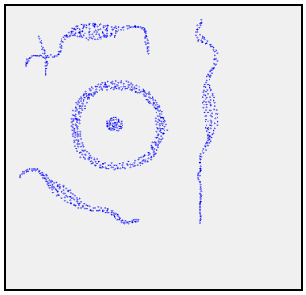
Fig.13 experiment 3

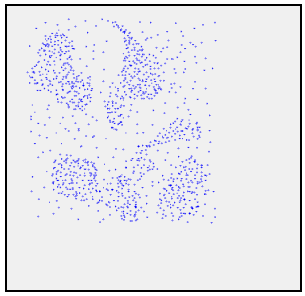
Fig.14 experiment 4

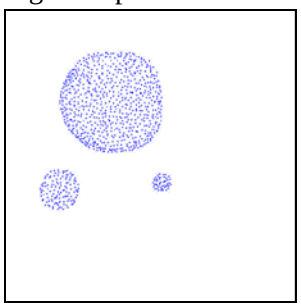
Fig.15 experiment 5


Fig.16 experiment 6

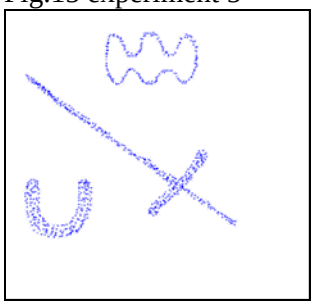
Fig.17 experiment 7

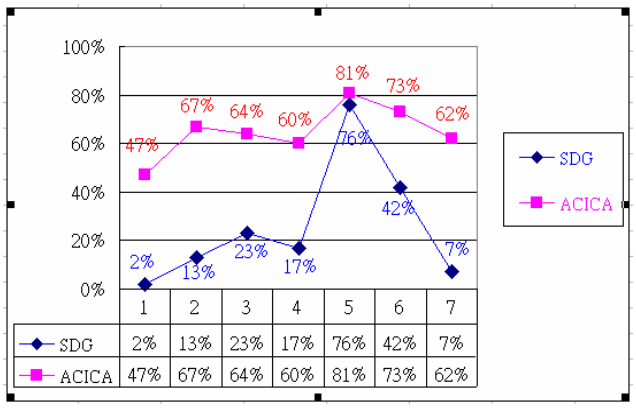

Fig.18 correct rates of ACICA and SDG

rates comparison is by using random 100 sets of parameters (density threshold, number of dividing parts in each dimension) from (16, 1) to (55, 3).

### 4.2 Discussion

In the ACICA algorithm, the density of each cell is calculated. When the total number of data is n and each dimension, total d dimensions, is divided into m intervals, there will be $m^d$ cells. The time of allocating the data and checking the density of all cells is k0*n. If p(=2d) is the number of nearby cells of one cell, the time of checking if the cell significant is $k1*p*[m^d + (m+1)^d]$ at most. The time of clustering in ADCC is $k2*[m^d + (m+1)^d + r]$, where r is the number of rules used in revised function $CR()$, but the time of clustering in ACICA is only $k3*(m^d + s) < k2*[m^d + (m+1)^d + r]$, where s is the number of significant cells in second grid structure and $r << s << m^d$. In the end, the time of checking the cluster's number of all data is k4*n. So, the time of clustering by using ACICA is shorter than using ADCC.

## 5. Conclusion and Future Work

In this paper, we propose the Adaptive Crossover-Imaged Clustering Algorithm (ACICA), which makes important contribution to support the obvious wider ranges of size of the cell and threshold of the density to reduce the drawbacks of grid-based clustering algorithms. We also discuss the details and advantages of ACICA to compare to ADCC, which is the first grid-based algorithm to support the obvious wider ranges of size of the cell and threshold of the density. And the clustering results of ACICA are exactly the same as the results of ADCC. From the present results obtained, it is fast and simple to realize that the ACICA algorithm not only inherits the advantage with the low time complexity, but also is verified by experiment that it outperforms than ADCC.

*References:*


[1] J. MacQieen. Some methods for classification and analysis of multivariate observation. *Proc. 5th Berkeley Symp. Math. Statist, Prob.*, 1:281-297,1967

[2] L. Kaufman and P.J. Rousseeuw. Finding Groups in Data: *An Introduction to Cluster Analysis*. New York: John Wiley & Sons, 1990.

[3]Charu C. Aggarwal, Philip S. Yu, "An effective and efficient algorithm for high-dimensional outlier detection" *The VLDB journal*, 14:211-221, 2005

[4] M. Ester, H. Kriegel, J. Sander, and X. Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *In Proc. of 2nd Int. Conf. on KDD*, 1996, pages 226-231.

[5] A. Hinneburg and D. A. Keim,. "An Efficient Approach to Clustering in Large Multimedia Databases with Noise", *In Knowledge Discovery and Data Mining*, 1998, pages 58-65.

[6] ANKERST M. etc. "OPTICS: Ordering Points to Identify the Clustering Structure." In Proc. ACM SIGMOD Int. Conf. on MOD, 1999, pages 49-60.

[7] A. H. Pilevar, M. Sukumar, "GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases", Pattern Recognition Letters 26(2005), 999-1010

[8] ZHAO Y.C., SONG J.,"GDILC: A Grid-based Density-Isoline Clustering Algorithm.", *In Proc. Internet. Conf. on Info-net*, Vol 3, pp.140-145,2001 ,

[9]Ma, W.M., Eden, Chow, Tommy, W.S., "A new shifting grid clustering algorithm", *Pattern Recognition* 37 (3),2004,503-514

[10]Alevizos, P., Boutsinas, B., Tasoulis, D., Vrahatis, M.N.,"Improving the K-windows clustering algorithm", *In Proc. 14th IEEE Internat. Conf. on Tools with Artificial Intell*, pp.239-245, 2002.

[11] Wang, Yang, R. Muntz, Wei Wang and Jiong Yang and Richard R. Muntz "STING: A Statistical Information Grid Approach to Spatial Data Mining", *In Proc. of 23rd Int. Conf. on VLDB*, 1997, pages 186-195.

[12] G. Sheikholeslami, S. Chatterjee, and A. Zhang. "WaveCluster: a wavelet-based clustering approach for spatial data in very large databases", *In VLDB Journal: Very Large Data Bases*, 2000, pages 289-304.

[13] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. "Automatic sub-space clustering of high dimensional data for data mining applications", *In Proc. of ACM SIGMOD Int. Conf. MOD*, 1998, pages 94-105.

[14] N. Lin, C. Chang, and C. Pan. "An Adaptive Deflect and Conquer Clustering algorithm", *In Proc. of 6th WSEAS Int. Conf. ACOS'07*, 2007, pages 156-160.