Rep. Fac. Sci. Engrg.
Saga Univ.
36-1 (2007)､33-38

Reports of the Faculty of Science and Engineering,
Saga University, Vol. 36, No.1, 2007

Automatic shape independent clustering with global optimum cluster determinations

By
Kohei Arai* and Ali Ridho Barakbah**

Abstract: A new method which allows identifying any shape of cluster patterns in case of numerical clustering is proposed. The method is based on the iterative clustering construction utilizing a nearest neighbor distance between clusters to merge. The method differs from other techniques of which the cluster density is determined based on calculating the variance factors. The cluster density proposed here is, on the other hand, determined with a total distance within cluster that derived from a total distance of merged cluster and the distance between merged clusters in the previous stage of cluster construction. Thus, the whole density for each stage can be determined by a calculated average of a total density within cluster of each cluster, and then split by referring the maximum furthest distance between clusters at that stage. Beside this, this paper also proposes a technique for finding a global optimum of cluster construction. Experimental results show how effective the proposed clustering method is for a complicated shape of the cluster structure

Key words: Single linkage hierarchical clustering method; Cluster density; Shape independent clustering; Automatic clustering

1. Introduction

For many years, many clustering algorithms have been proposed and widely used. It can be divided into two categories, hierarchical and non-hierarchical methods. It is commonly used in many fields, such as data mining, pattern recognition, image classification, biological sciences, marketing, city-planning, document retrieval, etc. The clustering means a process to define a mapping $f{:}D \rightarrow C$ from some data $D=\{d_1, d_2,\ldots,d_n\}$ to some clusters $C=\{c_1, c_2,\ldots, c_n\}$ based on similarity between d_i.

The task of finding a good cluster is very critical issues in clustering. Cluster analysis constructs good clusters when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external homogeneity) [2,8].

In fact, most authors find difficulty in describing clustering without some grouping criteria. For example, the objects are clustered or grouped on the basis of maximizing the inter-cluster similarity and minimizing the intra-cluster similarity [8]. One of the methods to define a good cluster is variance constraint [6] that calculates the cluster density with variance within cluster (v_w) and variance between clusters (v_b) [3,12]. The ideal cluster, in this case, has minimum v_w to express internal homogeneity and maximum v_b to express external homogeneity.

The parameter of v_w and v_b, however, can just be applied for identifying condensed clustering cases, which are the cluster members gathered in surrounding values so that the centroid resides in the circle weight of the members. Therefore, v_w and v_b can not be used in shape independent clustering, such as convex shape clustering.

One of the most famous clustering methods is hierarchical clustering. In hierarchical clustering the data are not partitioned into a particular cluster in at the first step. It runs with making a single cluster that has similarity, and then continues iteratively.

Hierarchical clustering algorithms can be either agglomerative or divisive [4,9,11]. Agglomerative method proceeds by series of fusions of the “n” similar objects into groups, and divisive method, which separate “n” objects successively into finer groupings. Agglomerative techniques are more commonly used.

One of similarity factors between objects in hierarchical methods is a single link that similarity closely related to the smallest distance between objects [1]. Therefore, it is called single linkage clustering method. Euclidian distance is commonly used to calculate the distance in case of numerical data sets [11]. For two-dimensional dataset, it performs as:

Received on Apr.28 2007

*Department of Information Science
**EEPIS:Electric Engineering Politechnique in Surabaya

$$d(x,y) = \sqrt{\sum_{i=1}^{n} | x_i - y_i |^2} \qquad (1)$$

The algorithm of single linkage clustering method is composed of the following steps:

1. Begin with an assumption that every point “n” is it’s own cluster c_i, where i=1..n.
2. Find the nearest distance between $m(c_r)$ and $m(c_u)$, where $r \neq u$ and $m(c_j)$ is members of cluster c_j.
3. Merge c_r and c_u into new cluster c_a where $m(c_a)$ is members fusion of c_r and c_u.
4. Repeat until it reaches an optimum

2. Cluster structure

In this section we describe the classification of cluster structure, which is divided into condensed cluster and shape independent cluster.

2. 1. Condensed cluster

The condensed cluster is defined as the cluster members gathered in closely surrounding locations as is shown in Fig.1. In the case of condensed cluster, the center of gravity resides in circle weight of the cluster members. The cluster density can, therefore, be determined with a calculated variance within cluster and variance between clusters.

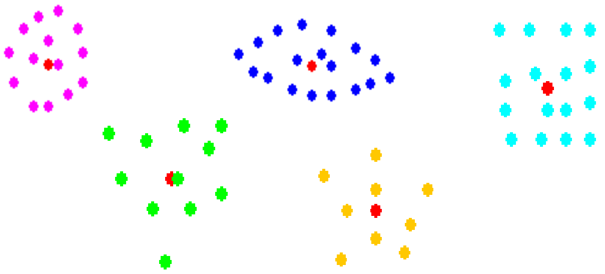

Fig.1 Examples of condensed clusters (centroids are shown with red dots).

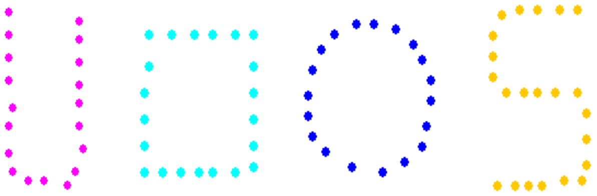

Fig.2 Examples of shape independent clusters

2.2. Shape independent cluster

The condensed cluster is very different from the shape independent cluster [10] that its similarity can be seen as such shape patterns. It, in this case, is very difficult to determine the centroid as is illustrated in Fig.2. Hence, the cluster density can not be defined by variance constraints.

3. Proposed algorithm

In this paper, a new simple algorithm of numerical clustering is proposed, in particular, for a shape independent clustering. The proposed method can also be applied in the case of condensed clustering. The grand idea of the proposed is utilizing single linkage hierarchical clustering method (SLHM) as a basic method to solve case of shape independent clustering. It is good method to get a hierarchy instead of an amorphous collection of groups [5]. The method itself actually can, however, not be applied for the case. To make SLHM as considerable method to solve the aforementioned case, there are two critical items those we must redefine: the cluster density and the global optimum. The proposed method can solve any cases of shape independent clustering. The problem solving by the proposed method is absolutely constant, even though the shape independent case is more complex. It is somewhat different from CLUSTER method [10] that is based on Relative Neighborhood Graph (RNG). The CLUSTER method needs to determine the appropriate threshold value more precisely and the suitable small discretization of the fraction Max-Min to solve the more complex shape independent cases.

3.1 Determining the cluster density

In this paper, a new clustering method with a definition of cluster density is proposed. It is very unique and simple, because we just utilize the total distance between clusters at each stage of cluster construction. Before calculating the cluster density, we calculate at first $d(i,j)$ as the nearest distance between clusters for each clusters j in stage i. Then, we set the minimum of $d(i,j)$ as below:

$$f(i) = \min(d(i,j)) \qquad (1)$$

If $и(a)=b$ expresses the nearest cluster a is b, then we look for the cluster characteristics will be merged that classified as:

1. $d(i,j)$ equals to $f(i)$
2. $и(a)=b$, $и(b)=a$, and $и(x_p)=\{a|b\}$ where $и(x_p)$ are clusters those their nearest cluster refer to a or b, and $p=1,\ldots,n$.

Therefore, if new cluster j constructed and i now is next stage, then the total distance within cluster $\sigma(i,j)$ for new cluster j in the next stage i can be defined as:

$$\sigma(i,j) = \sigma(i\text{-}1,a) + \sigma(i\text{-}1,b) + \sum_{p=1}^{n} \sigma(i\text{-}1,p) + f(i\text{-}1) \times (n+1) \qquad (2)$$

If $p(i,j)$ is number of members within cluster j in stage i, then the density within cluster $\delta(i,j)$ in cluster j in stage i can be defined as:

$$\delta\ (i,j) \frac{\sigma(i,j)}{p(i,j)} \qquad (3)$$

where j=1,2,…,number of cluster in stage i. After that, we calculate the average of $\delta\ (i,j)$. The last, we can calculate the density of all cluster δ_i in stage i as follows:

$$\delta_i = \frac{\overline{\delta(i,j)}}{f(i,j)} \qquad (4)$$

3.2. Finding the global optimum

The proposed technique is derived from analyzing the values moving pattern of δ at each stage. Then we identify the considerable formula to find the global optimum. After applying it to some experiments so that we can analyze the accuracy, we find the global optimum can be reached at the stage of i that has maximum ∂, where:

$$\partial_i = \delta_{i+1} \text{ - } \delta_i \qquad (5)$$

In order to construct cluster automatically, we put the threshold value λ to get a maximum ∂. The value of λ expresses the value of ∂×100. It means, if we set λ=1, the global optimum can be reached at the stage of i when $\partial_i > 0.01$. The more amorphous shape independent case needs smaller λ to set as more precise as possible. By setting the value of λ, the well-separated cluster will be constructed.

4. Experimental results

We examine our proposed algorithm to some of different cases for the shape independent clustering. It covers determining the cluster density as well as the global optimum. Various cases those examined can determine the accuracy of the proposed method. For every case, we record the valuable data that has moving values of each stage. We also involve the accuracy parameter to observe the performance of all clustering cases.
In our experimental cases, we use λ=1-2 to reach the global optimum. We also use an additional variable φ to express the different values between max(∂) and ∂_i that has candidate max(∂).

$$\varphi = \frac{\max(\partial)}{\text{Candidate max}(\partial)} \qquad (6)$$

The value of φ can show a distant value to get global optimum. The high φ, at least $\varphi \geq 2$, expresses possibility to construct well-separated clusters. It avoids cluster construction reaching any local optima. If the closer value is non-positive, the value of φ will be ϕ, means the global optimum is absolutely right and the well separations between clusters are constructed very well. The value of ϕ,can represent the performance of cluster construction and precision of the proposed method to solve the clustering cases. Fig.3 and 4 shows how the proposed method works for avoidance of the local minima in comparison to the U-shape data set as is shown in Fig. 3.

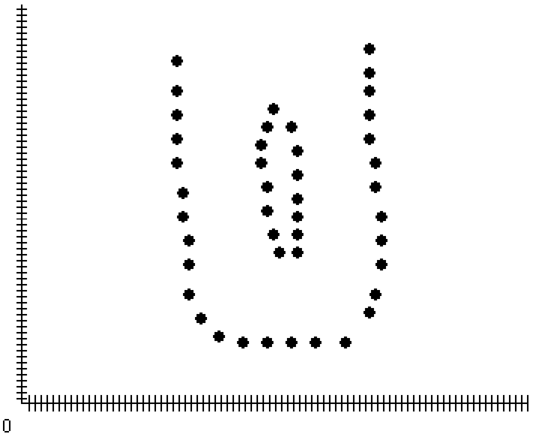

Fig. 3. U-shape data set

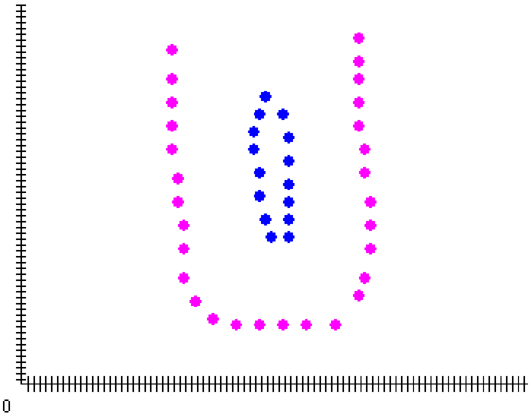

Fig. 4. The result of applying the proposed method

It is found that our proposed method is superior to the existing shape independent clustering in this clustering case. The result shows the accuracy of ∂ to express the global optimum, as viewed in Fig. 5. In the top case of Fig.5, n=44. We use λ=2 to reach the global optimum. The experimental result showed that the maximum of ∂= 0.03228, in the stage 4 which numbers of well-separated cluster is 2. It is proved that the global optimum will be reached with 2 numbers of clusters. The value φ = 6.0757. In the bottom case of Fig.5, n=36. We use λ=1 to reach the global optimum. The experimental result showed that the maximum of ∂ = 0.026139, in the stage which numbers of well-separated cluster is 2. It is proved that the global optimum will be reached with 2 numbers of clusters. The value φ = 23.5486.

We applied the proposed method to solve some various shape independent cases (Fig. 6 - Fig. 12). The result of clustering construction is indicated with a different color.

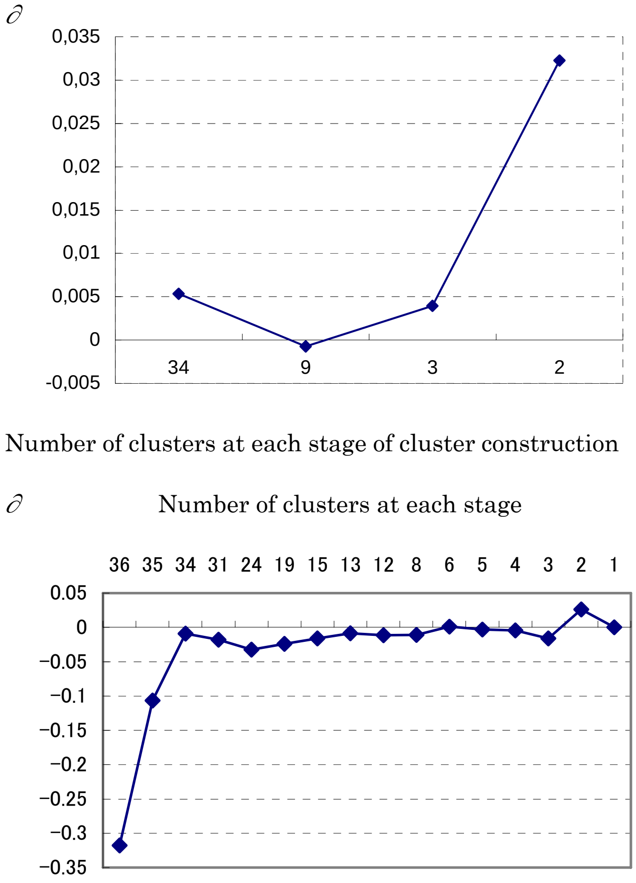

Fig. 5. The moving values of ∂ at each stage

From the experiment, we can see that the proposed algorithm computed high φ. Actually, the high φ expresses the high possibility to construct well-separated clusters. The limitation of φ that we set is ≥ 2.It means if $\varphi < 2$, there is another ∂ which is very close to be global optimum that we determine as max(∂).

The experiment results with several data sets perform high φ. Moreover, some of data sets, interrelated and complex interrelated data set, reach ϕ. We can also see that our proposed algorithm can solve the more complex shape independent clustering cases, such as Nested diamond data set. In the nested diamond data set, the accuracy performs high φ=5.7585.

Table 1 perform the gap values between max(∂) and candidate max(∂) for all data sets.

Table 1 Gap value between max(∂) and candidate max(∂)

Data set	Gap value (×100)
Interrelated	4.0855
S-shape	4.9559
Nested circle	9.5790
Contiguous	6.5495
Nested diamond	7.3309
Complex interrelated	1.9269
Random	0.8617

It is found that our proposed method is superior to the existing shape independent clustering in this clustering case. The result shows the accuracy of ∂ to express the global optimum, as viewed in Fig. 5. In this case, n=44. We use λ=2 to reach the global optimum.

The experimental result showed that the maximum of ∂= 0.03228, in the stage 4 which numbers of well-separated cluster is 2. It is proved that the global optimum will be reached with 2 numbers of clusters. The value φ = 6.0757.

We applied the proposed method to solve some various shape independent cases (Fig.6-Fig. 12). The result of clustering construction is indicated with a different color..

Besides applying to the shape independent clustering cases, we also tried to apply our proposed method in condensed clustering case (Fig.13). The result showed the usability of the proposed method to apply in condensed clustering cases.

The use of threshold is very simple. In most of cases, we usually set λ=2. But, for an amorphous shape independent clustering cases, such as complex interrelated and random data set, we need to set λ smaller. In those cases, we set λ=1.

Besides applying to the shape independent clustering cases, we also tried to apply our proposed method in condensed clustering case (Fig.13).

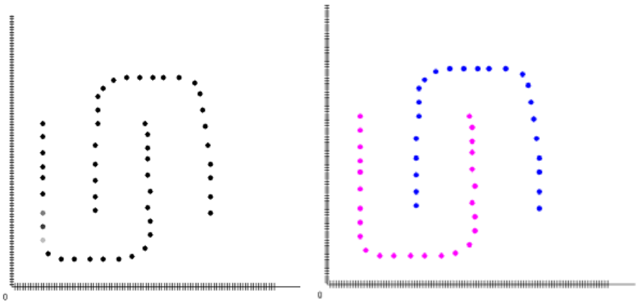

Fig. 6. Interrelated data set, $n=50$, $\lambda=1$, $\max(\partial) = 0.020407$, $\varphi = \phi$.

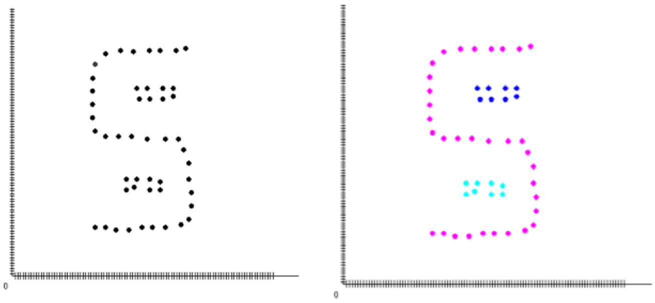

Fig. 7. S shape data set, $n=49$, $\lambda=1$, $\max(\partial) = 0.026284$, $\varphi = 3.647$.

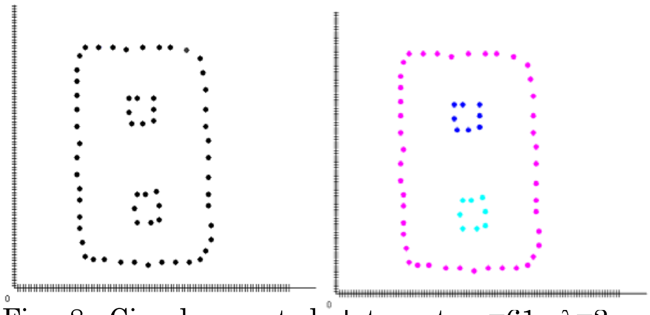

Fig. 8. Circular nested data set, $n=61$, $\lambda=2$, $\max(\partial) = 0.05081$, $\varphi = 3.9418$.

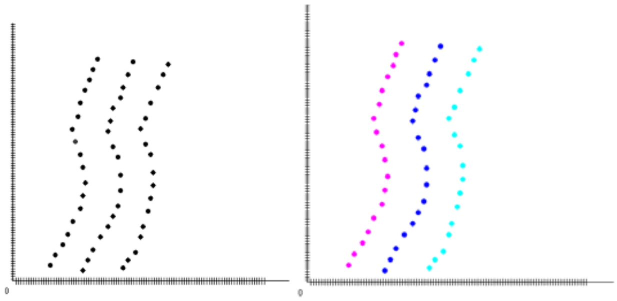

Fig. 9. Contiguous data set, $n=53$, $\lambda=1$, $\max(\partial) = 0.015496$, $\varphi = \phi$.

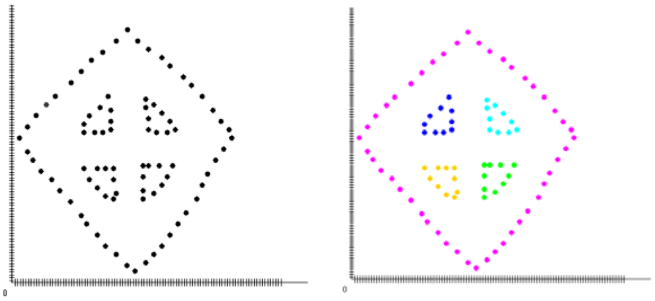

Fig. 10. Diamond nested data set, $n=76$, $\lambda=1$, $\max(\partial) = 0.035718$, $\varphi = 5.6072$.

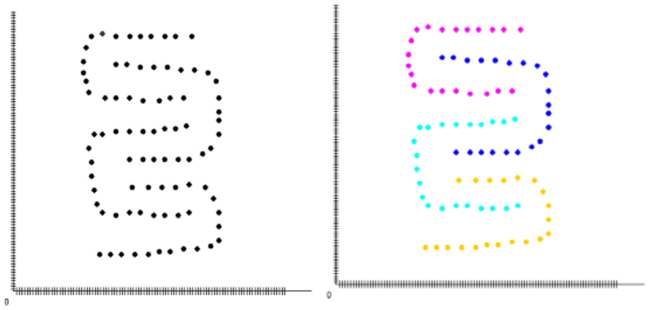

Fig. 11. Complex interrelated data set, $n=84$, $\lambda=1$, $\max(\partial) = 0.014641$, $\varphi = 16.16$.

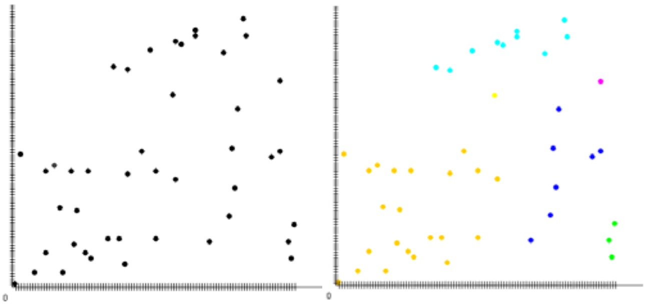

Fig. 12. Random data set, $n=43$, $\lambda=0.5$, $\max(\partial) = 0.009148$, $\varphi = 4.3458$.

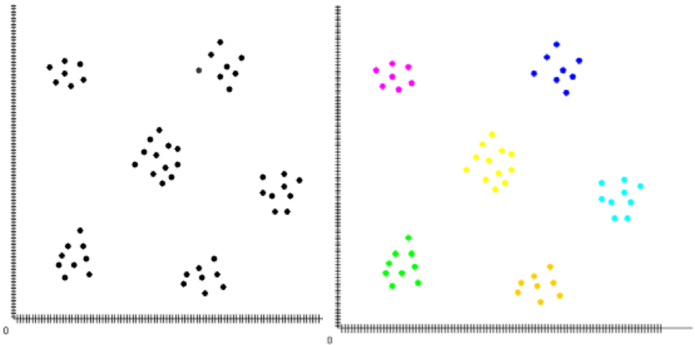

Fig. 13. Normal data set, $n=43$, $\lambda=1$, $\max(\partial) = 0.232249$, $\varphi = 49.7534$.

The high φ =25.4074 expressed that the proposed algorithm is able to make automatic well-separated clusters. Besides the gap value that computed is very large, 16.5165 (×100). The result showed the usability of the proposed method to apply in condensed clustering cases.

5. Conclusions

It is found that the proposed method can be used for shape independent clustering as well as condensed clustering. From the experimental results with some various clustering cases, the proposed method can solve the clustering problem and create well-separated clusters.

The variable φ showed in those cases that the possibility of constructing well-separated clusters is high, implies that the proposed method can also avoid any local optima and find the global optimum.

The threshold of λ is easy to set ensuring reach the global optimum. For more the amorphous shape independent cases need smaller λ to set as more precise as possible. By setting the value of λ, the well-separated cluster will be constructed. The very high value of φ for normal data sets proves that the proposed method is also considerable to solve the problems for condensed clustering cases.

References

[1] G. Karypis, E.H. Han, V. Kumar, Chameleon: a Hierarchical Clustering Algorithm using Dynamic Modeling, IEEE Computer: Special Issue on Data Analysis and Mining 32(8):68W5, 1999.

[2] G.A. Growe, Comparing Algorithms and Clustering Data: Components of The Data Mining Process, thesis, department of Computer Science and Information Systems, Grand Valley State University, 1999.

[3] S. Ray and R.H. Turi, “Determination of Number of Clusters in K-means Clustering and Application in Colour Image Segmentation”, 4th ICAPRDT Proc., pp.137-143, 1999.

[4] M. Halkidi, Y. Batistakis, M. Vazirgiannis, “Clustering Algorithms and Validity Measures”, Proceedings of The 13th International Conference on Scientific and Statistical Database Management, July 18–20. IEEE Computer Society, George Mason University, Fairfax, Virginia, USA, 2001.

[5] A.W. Andrew, K-means and Hierarchical Clustering, School of Computer Science, Carnegie Mellon University, 2001.

[6] C.J. Veenman, M.J.T. Reinders, and E. Backer, “A Maximum Variance Cluster Algorithm”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 9, pp. 1273-1280, September, 2002.

[7] M. Gaertler, Clustering with Spectral Methods, thesis, Universitat Konstanz, Fachbereich Mathematik und Statistik, Fachbereich Informatik und Informationswissenschaft, 2002.

[8] V. Estivill-Castro, “Why So Many Clustering Algorithms-A Position Paper”, ACM SIGKDD Explorations Newsletter, Volume 4, Issue 1, pp. 65-75, 2002.

[9] D. Frossyniotis, A. Likas and A. Stafylopatis, “A Clustering Method Based on Boosting”, Pattern Recognition Letters (2004) .

[10] S. Bandyopadhyay, An Automatic Shape Independent Clustering Technique, Machine Intelligence Unit, Journal of Pattern Recognition Society, volume 37, number 1, January 2004.

[11] P.A. Vijaya, M.N. Murty, and D.K. Subramanian, Leaders–Subleaders: An Efficient Hierarchical Clustering Algorithm for Large Data Sets, Pattern Recognition Letters 25 (2004) 505–513.

[12] W.H. Ming and C.J. Hou, Cluster Analysis and Visualization, Workshop on Statistics and Machine Learning, Institute of Statistical Science, Academia Sinica, 2004.