Kohei Arai, Ali Ridho Barakbah, "Method for shape independent clustering in case of numerical clustering together with condensed clustering",
The 8th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2004), July 18-21, 2004, Orlando, Florida.

The 8th World Multiconference on Systemics, Cybernetics and Informatics
July 18 - 21, 2004 Orlando, Florida, USA

Method for Shape Independent Clustering in case of Numerical Clustering together with Condensed Clustering

Kohei ARAI
Information Science, Saga University,
Saga, Japan

and

Ali Ridho BARAKBAH
Information Science, Saga University,
Saga, Japan

ABSTRACT

A new method which allows to identify 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.

Keywords: Clustering, Single Linkage Hierarchical clustering method, Cluster density, and Shape independent clustering.

1. INTRODUCTION

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 suggestions for grouping criteria. For example, "the objects are clustered or grouped based on the principles of maximizing the inter-class similarity and minimizing the intra-class 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 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 t_i gathered in surrounding values so that the centroid resides in the circle weight of t_i. Therefore, V_w and V_b can not be used in shape independent clustering cases, such as convex shape clustering.

2. 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 clustermembers.The cluster density can, therefore, be determined with a calculated variance within cluster and variance between clusters.

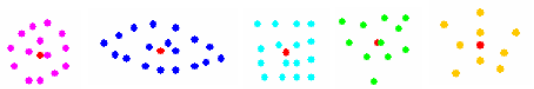

Fig. 1. Examples of condensed clusters which center of gravity (red color) is in the circle weight of cluster members.

3. 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. Hence, the cluster density can not be defined by variance constraints.

Fig. 2. Examples of shapes of the independent cluster

4. THE PROPOSED METHOD

In this paper, a new simple method 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

Kohei Arai, Ali Ridho Barakbah, "Method for shape independent clustering in case of numerical clustering together with condensed clustering",
The 8th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2004), July 18-21, 2004, Orlando, Florida.

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.

5. DETERMINING OF 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:

$$f(i) = \min(d(i,j)) \tag{1}$$

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(Xp)={a|b} where и(Xp) are clusters those their nearest cluster refer to a or b, and p=1..n.

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

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

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

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

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

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

6. FINDING 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:

$$\partial_i = \delta_{i+1} - \delta_i \tag{5}$$

In order to construct cluster automatically, we put the threshold value λ to get a maximum ∂. The value of λ expresses the value of ∂x100. It means, if we set λ=0.5, the global optimum can be reached at the stage of i when ∂_i>0.005. 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.

9. ACCURACY OF THE PROPOSED METHOD

We examined our proposed method to some of different cases for the shape independent clustering. It covered 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 values moving pattern at each stage. In our experimental cases, we use λ=0.5-2 to reach the global optimum. We also use an additional variable φ to express the different values between max(∂) and ∂_i that has closer value to max(∂).

$$\varphi = \frac{\max(\partial)}{\text{closer value to } \max(\partial)} \tag{6}$$

The value of φ can show a distant value to get global optimum. The large φ, at least φ≥2, expresses possibility to construct well-separated cluster. 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.

Fig.4 shows how the proposed method works for avoidance of the local minima in comparison to the existing shape independent clustering as is shown in Fig. 3.

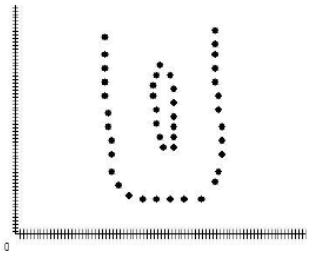

Fig. 3. The shape independent clustering case

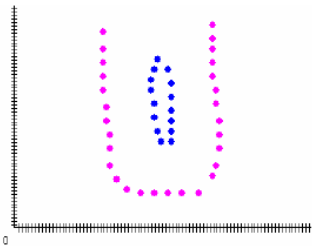

Fig. 4. The result of applying the proposed method

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 this case, n=44. We use λ=1 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.

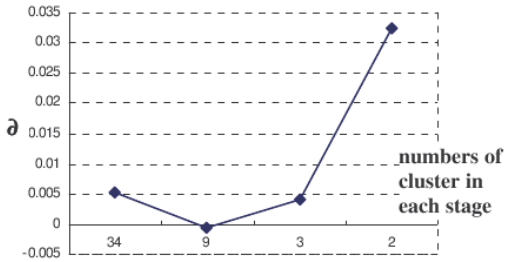

Fig. 5. The moving values of ∂ at each stage

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.

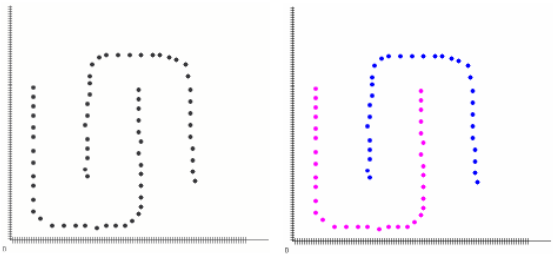

Fig. 6. Interrelated data set, n=67, λ=1,
max(∂) = 0.037249, φ = ϕ.

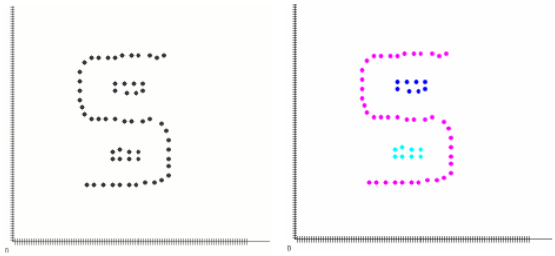

Fig. 7. S-shape data set, n=58, λ=2, max(∂) = 0.063371,
φ = 4.5881.

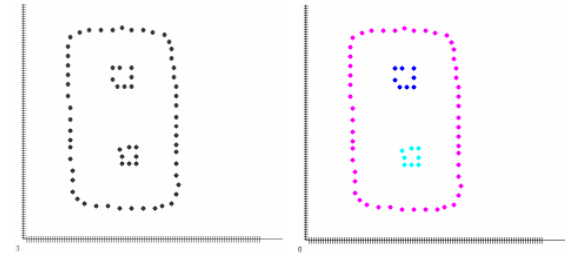

Fig. 8. Circular nested data set, n=68, λ=2,
max(∂) = 0.107077, φ = 9.48675.

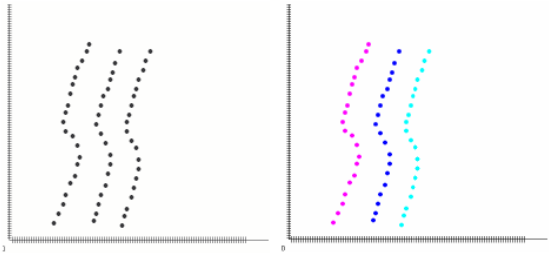

Fig. 9. Contiguous data set, n=60, λ=1,
max(∂) = 0.073639, φ = 9.04212.

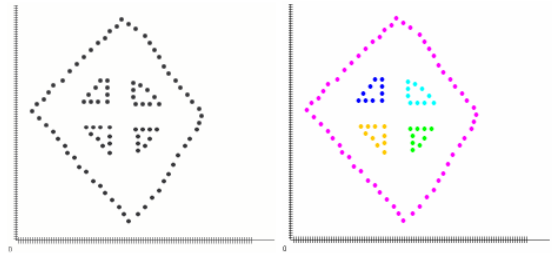

Fig. 10. Diamond nested data set, n=96, λ=2,
max(∂) = 0.088715, φ = 5.7585.

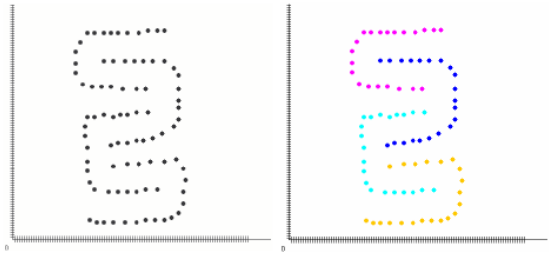

Fig. 11. Complex interrelated data set, n=84, λ=0.5,
max(∂) = 0.018349, φ = ϕ.

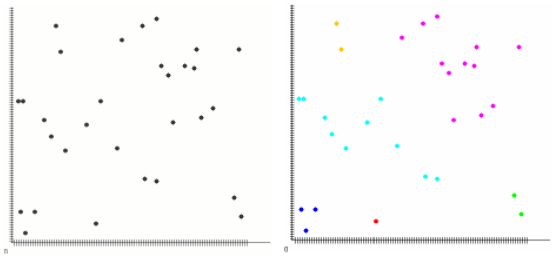

Fig. 12. Random data set, n=30, λ=1, max(∂) = 0.0014768, φ = 2.4009.

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.

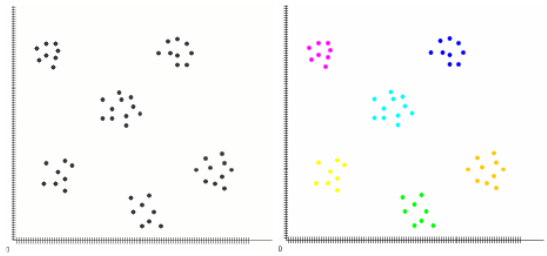

Fig. 13. Normal data set, n=54, λ=1, max(∂) = 0.171932, φ = 25.4074.

The total stage of cluster construction for each data set figured at Table. 1.

Data set	Total stage
Interrelated	4
S-shape	3
Circular nested	3
Contiguous	3
Diamond nested	3
Complex interrelated	4
Random	14
Normal	4

Table 1. Total stage to get optimal cluster for each data set

10. CONCLUSION

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, 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.

11. 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)** (Accepted).

[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.