Comparative study between the proposed shape independent clustering method and the conventional methods (K-means and the other)

Kohei Arai [1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

Cahya Rahmad [2]
Electronic Engineering Department
The State Polytechnics of Malang,
East Java, Indonesia

***Abstract*—Cluster analysis aims at identifying groups of similar objects and, therefore helps to discover distribution of patterns and interesting correlations in the data sets. In this paper, we propose to provide a consistent partitioning of a dataset which allows identifying any shape of cluster patterns in case of numerical clustering, convex or non-convex. The method is based on layered structure representation that be obtained from measurement distance and angle of numerical data to the centroid data and based on the iterative clustering construction utilizing a nearest neighbor distance between clusters to merge. Encourage result show the effectiveness of the proposed technique.**

Keywords—clustering algorithms; mlccd; shape independence clustering;

I. INTRODUCTION

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data and is also broadly recognized as a useful tool in many applications. It has been subject of wide research since it arises in many application domains in engineering, business and social sciences. Researchers of many disciplines have addressed the clustering problem. The objective of algorithms is to minimize the distance of the objects within a cluster from the representative point of this cluster[1]. The clustering means process to define a mapping, f: D → C from some data D={t1, t2,t3,…tn} to some clusters C={c1, c2,c3,…, cn} based on similarity between ti.

Clustering is a central task for which many algorithms have been proposed. The task of finding a good cluster is very critical issues in clustering[2]. Cluster analysis constructs good clusters when the members of a cluster have minimize distances (Intra-cluster distances are minimized or internal homogeneity) are also not like members of other clusters (Inter-cluster distances are maximized). Clustering algorithms can be hierarchical or partitional[3].

In hierarchical clustering, the output is a tree showing a sequence of clustering with each cluster being a partition of the data set. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones. The result of the algorithm is a tree of cluster, called dendogram, which shows how the cluster are related, by cutting the dendogram at a desired level, a clustering of data the data items into disjoint groups is obtained [1]. The most well known methods for clustering is K-means developed by Mac Queen in 1967. K-means is a partition clustering method that separates data into k mutually excessive groups. Partitional clustering attempts to directly decompose the data set into a set of disjoint clusters. By iterative such partitioning, K-means minimizes the sum of distance from each data to its clusters. This method ability to cluster huge data, and also outliers, quickly and efficiently[4]. However, K-means algorithm is very sensitive in initial starting points. Because of initial starting points generated randomly, K-means does not guarantee the unique clustering results. However, many of above clustering methods require some additional user specified parameters, such as shape of cluster, optimal number, similirity threshold etc. In this paper, a new simple algorithm of numerical clustering is proposed. The proposed method, in particular, for a shape independent clustering that can also be applied in the case of condensed clustering.

II. PROPOSED CLUSTERING ALGORITHM

In this paper, a new simple algorithm of numerical clustering is proposed. The proposed method, in particular, is for a shape independent clustering. The proposed method can also be applied in the case of condensed clustering. we implemented multi-layer centroid contour distance (mlccd)[5] with some modification to the clustering.

The algorithm is described as follow:

1) Begin with an assumption that every point "n" is it's own cluster c_i, where i =1,2,…n

2) Calculate the centroid location

3) Calculate the angle and distance for every point to the centroid

4) Make multi layer centroid contour distance based on step 2 and 3

5) Set i = 1 as initial counter

6) Increment i = i + 1

7) Measure distance between cluster in the location i with cluster in the location i-1 and cluster in the location i+1.

8) Merge two cluster become one cluster base on the a nearest neighbor distance between clusters (see equation 2.4) as shown in the step 7

9) Repeat from step 6 to step 8 while i< 360

10) Repeat step 5 to step 9 until the required criteria is met

Firstly in the step 1, let every point "n" is it's own cluster, if there are n data it mean there are n cluster. Secondly, in the step 2 obtain the centroid by using equation 2.1 in this case every point have contribution to find the centroid location. Step3 Calculate the distance and angle between centroid to the every point, the angle is obtained by using arctangent of dy/dx, dy is differences y position between position y of every point and position y in the centroid and dx is differences x position between position x of every point and position x in the centroid see equation 2.2. The distance between centroid to the every point of data can be obtained by using equation (2.3). The next step By using the angle of every point and its distance then make the multi layer centroid contour distance (mlccd). The next process set i as counter start from 1 to 360 then calculate distance by using Euclidian distance between every cluster that pointed by i and every cluster that pointed by i-1 and i+1. After the distance was calculated then merge two cluster become one cluster base on the nearest distance.

position of the centroid is:

$$X_c = \frac{X_1+X_2+X_3+....+X_n}{n} \quad , \quad Y_c = \frac{Y_1+Y_2+Y_3+....+Y_n}{n} \qquad (1)$$

Where:

X_c = position of the centroid in the x axis

Y_c = position of the centroid in the y axis

n = Total point or data (every point have x position and y position)

angle every point to the centroid is:

$$\text{Angle} = \text{arctangent}\left(\frac{dy}{dx}\right) \qquad (2)$$

Where:

$dx = x - x_c$

$dy = y - y_c$

(x_c, y_c) = Position centroid

(x,y) = Position point or data

The distance from the centroid to every point is :

$$\text{Dis} = \sqrt{dx^2 + dy^2} \qquad (3)$$

Where:

$dx = x - x_c$

$dy = y - y_c$

(x_c, y_c) = Position centroid

(x,y) = Position point or data

The a nearest neighbor distance between clusters is calculated by using Euclidian distance these equation is commonly used to calculate the distance in case of numerical data sets [6]. For two-dimensional dataset, it performs as:

$$d(A, B) = \sqrt{\sum_{i=1}^{n} |A_i - B_i|^2} \qquad (4)$$

III. Experiment Result

In order to analyze the accuracy of our proposed method we represent error percentage as performance measure in the experiment. It is calculated from number of misclassified patterns and the total number of patterns in the data sets[7] (see the equation 3.1). We compare our method with the conventional clustering methods (k-means and other) to the same dataset.

$$\text{Error} = \frac{number\ of\ misclassified\ patterns}{number\ of\ patterns}\ X\ 100\% \qquad (5)$$

We applied the proposed method to solve some various shape independent cases and also tried to apply the proposed method in condensed clustering case.The dataset consist Circular nested dataset contain 96 data and 3 cluster with cluster1 contain 8 data, cluster 2 contain 32 data and cluster 3 contain 56 data, inter related dataset contain 42 data and 2 cluster with cluster 1 contain 21 data and cluster 2 contain 21 data, S shape dataset contain 54 data and 3 cluster with cluster1 contain 6 data, cluster 2 contain 6 data and cluster 3 contain 42 data. u shape dataset contain 38 data and 2 cluster with cluster 1 contain 12 data and cluster 2 contain 26 data. The 2 cluster Random dataset contain 34 data and 2 cluster with cluster1 contain 15 data and cluster 2 contain 19. The 3 cluster condense dataset contain 47 data and 3 cluster with cluster1 contain 16 data, cluster 2 contain 14 data and cluster 3 contain 17. The last data is 4 cluster condense dataset contain 64 data and 4 cluster with cluster1 contain 14 data, cluster 2 contain 18 data cluster 3 contain 15 data and cluster 4 contain 17 data. In the Figure 1 up to Figure 7, A is unlabelled data, B is labelled data by using proposed method and C is labelled data by using K-mean method.

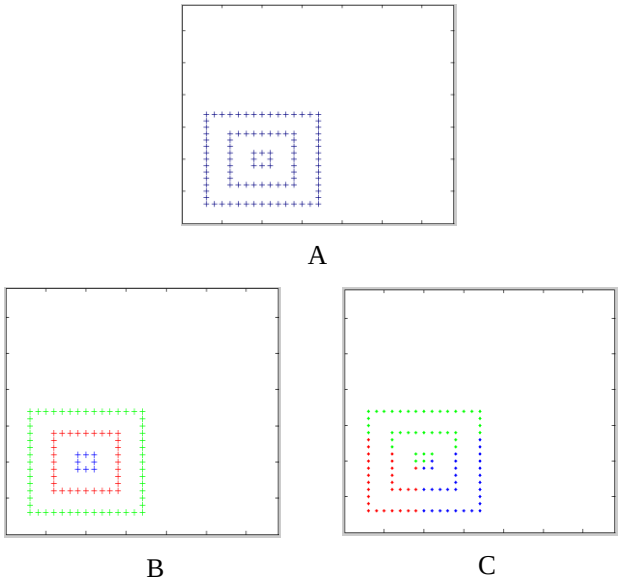

Fig.1. Circular nested dataset contain 96 data and 3 cluster with cluster1 contain 8 data, cluster 2 contain 32 data and cluster 3 contain 56 data

In Figure 1 by using proposed method there is no misclassified, by using k-mean there are misclassified in the cluster1 4 miss, cluster2 29 and cluster3 35miss, average error is 67.7%.

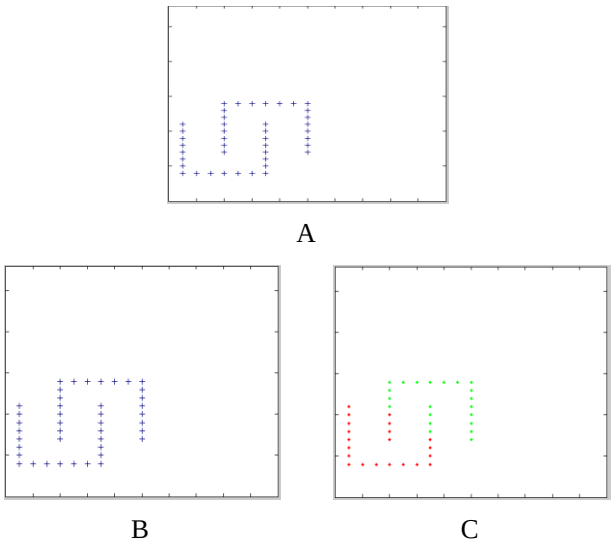

Fig.2. inter related dataset contain 42 data and 2 cluster with cluster 1 contain 21 data and cluster 2 contain 21 data

In figure 2 by using proposed method there is no misclassified, by using k-mean there are misclassified in the cluster1 4 miss and cluster2 4 miss, average error is 19.04%.

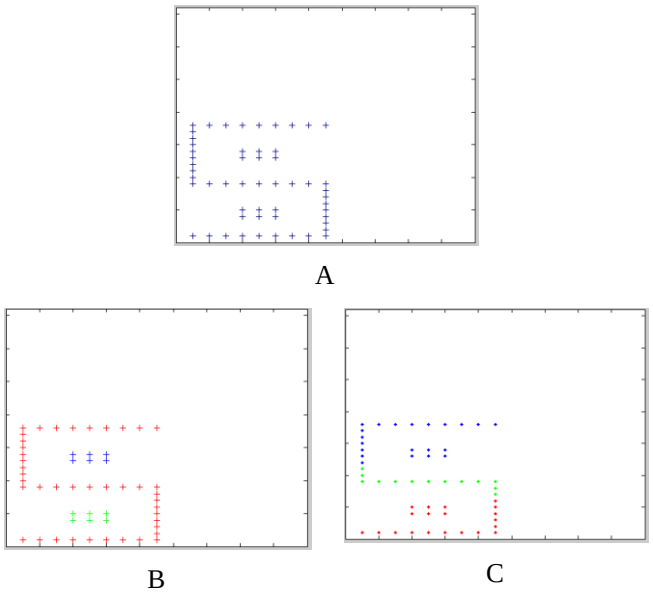

Fig.3. S shape dataset contain 54 data and 3 cluster with cluster1 contain 6 data, cluster 2 contain 6 data and cluster 3 contain 42 data

In Figure 3 by using proposed method there is no misclassified, by using k-mean there are misclassified in the cluster1 0 miss, cluster2 0 and cluster3 29miss, average error is 23.01%.

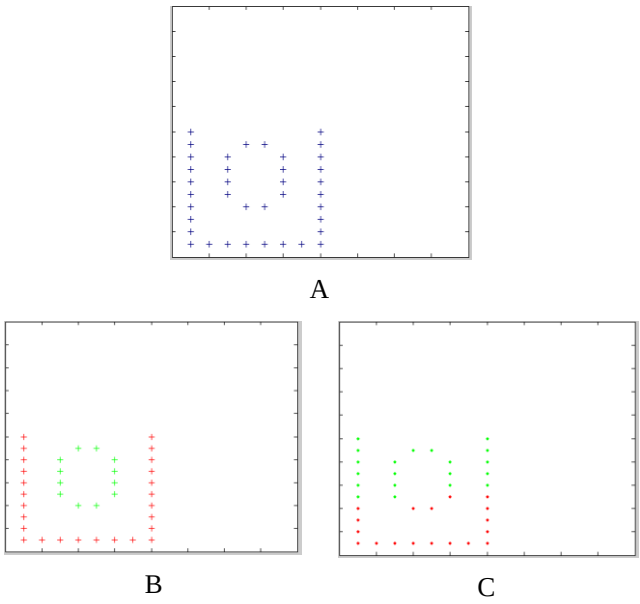

Fig.4. u shape dataset contain 38 data and 2 cluster with cluster 1 contain 12 data and cluster 2 contain 26 data.

In Figure 4 by using proposed method there is no misclassified, by using k-mean there are misclassified in the cluster1 3 miss and cluster2 11 miss, average error is 33.65%.

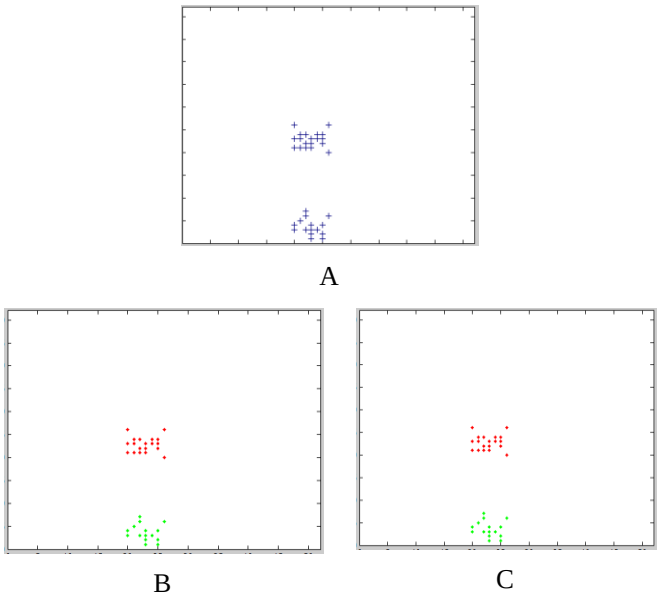

Fig.5. The 2 cluster Random dataset contain 34 data and 2 cluster with cluster1 contain 15 data and cluster 2 contain 19.

In Figure 5 by using proposed method and k-mean there is no misclassified average error is 0%.

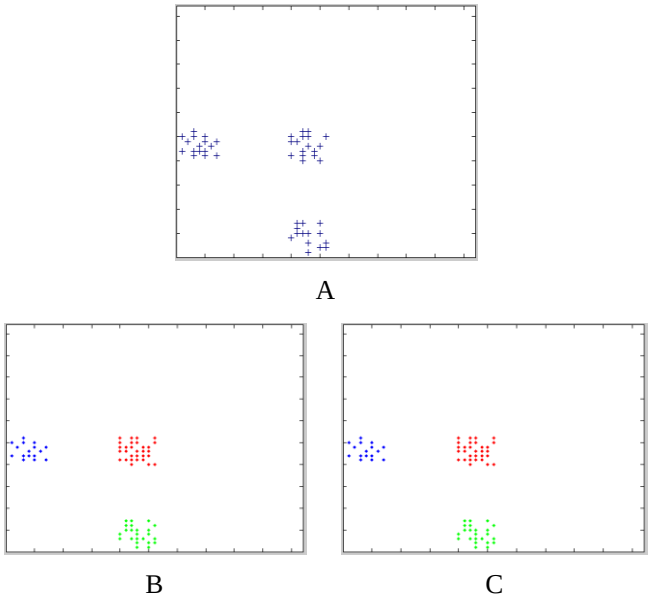

Fig.6. The 3 cluster condense dataset contain 47 data and 3 cluster with cluster1 contain 16 data, cluster 2 contain 14 data and cluster 3 contain 17

In Figure 6 and Figure 7 by using proposed method and k-mean there is no misclassified average error is 0%.

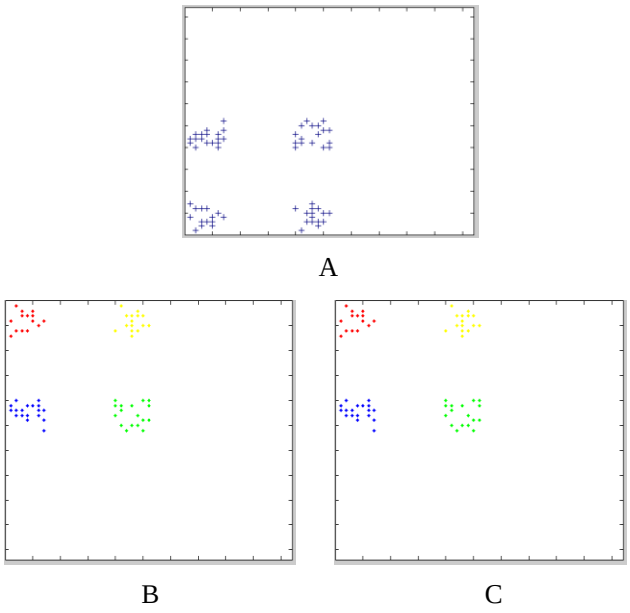

Fig.7. The 4 cluster condense dataset contain 64 data and 4 cluster with cluster1 contain 14 data, cluster 2 contain 18 data cluster 3 contain 15 data and cluster 4 contain 17 data.

We also implement some other clustering method that are Hierarchical algorithms (Single Linkage, Centroid Linkage, Complete Linkage and Average Linkage) to the same dataset the average result shown in the Table2.

TABLE I. AVERAGE ERROR IN PERCENT OF CLUSTERING BY USING LAYERED STRUCTURE REPRESENTATION AND CLUSTERING BY USING K-MEAN

Dataset	Propose method Error in %	k-mean Error in %
Circular nested	0	67.7
inter related	0	19.04
S shape	0	23.01
u shape	0	33.65
2 cluster condense	0	0
3 cluster condense	0	0
4 cluster condense	0	0
Average	0	20.48

TABLE II. AVERAGE ERROR IN PERCENT OF CLUSTERING BY USING HIERARCHICAL CLUSTERING

Clustering Method	Single linkage	Centroid linkage	Complete linkage	average linkage
Average	19.32	57.82	57.94	58.21

In the Table 1 and Table2 are shown that the proposed method compare with other method allows to identifying any shape of cluster as well as condensed dataset.

IV. CONCLUSION

In this work a new clustering methodology has been introduced based on the layered structure representation that be obtained from measurement distance and angle of numerical data to the centroid data and based on the iterative clustering construction utilizing a nearest neighbor distance between clusters to merge. From the experimental results with some various clustering cases, the proposed method can solve the clustering problem and create well-separated clusters. It is found that the proposed provide a consistent partitioning of a dataset which allows identifying any shape of cluster patterns in case of numerical clustering as well as condensed clustering.

ACKNOWLEDGEMENT

The authors would like to thank all laboratory members for their valuable discussions through this research.

REFERENCES

[1] M. Halkidi, "On Clustering Validation Techniques," *Journal of Intelligent Information Systems*, pp. 107–145, 2001.
[2] V. Estivill-Castro, "Why so many clustering algorithms: a position paper," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 1, pp. 65–75, 2002
[3] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.
[4] H. Ralambondrainy, "A conceptual version of the K-means algorithm," *Pattern Recognition Lett.*, pp. 1147–1157, 1995.
[5] K. Arai and C. Rahmad, "Content Based Image Retrieval by using Multi Layer Centroid Contour Distance," vol. 2, no. 3, pp. 16–20, 2013.
[6] P. A. Vijaya, M.N. Murty, and D. K. Subramanian, "An Efficient Hierarchical Clustering Algorithm for Large Data Sets," *Pattern Recognition Letters 25*, pp. 505–513, 2004.
[7] K. Arai and A. R. Barakbah, "Cluster construction method based on global optimum cluster determination with the newly defined moving variance," vol. 36, no. 1, pp. 9–15, 2007.

AUTHIORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission A of ICSU/COSPAR since 2008. He wrote 30 books and published 307 journal papers.

Cahya Rahmad, He received BS from Brawijaya University Indonesia in 1998 and MS degrees from Informatics engineering at Tenth of November Institute of Technology Surabaya Indonesia in 2005. He is a lecturer in The State Polytechnic of Malang Since 2005 also a doctoral student at Saga University Japan Since 2010. His interest researches are image processing, data mining and patterns recognition.