Genetic Algorithm Utilizing Image Clustering with Merge and Split Processes Which Allows Minimizing Fisher Distance Between Clusters

Kohei Arai [1]

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract—Genetic algorithm utilizing image clustering with merge and split processes which allows minimizing Fisher distance between clusters is proposed. Through experiments with simulation and real remote sensing satellite imagery data, it is found that the proposed clustering method is superior to the conventional k-means and ISODATA clustering methods in comparison to the geographic maps and classification results from Maximum Likelihood classification method.

Keyword—*image clustering; clustering; genetic algorithm; Fisher distance*

I. INTRODUCTION

There are hierarchical and non-hierarchical clustering methods. In particular, k-means and ISODATA clustering methods are well known as conventional clustering methods with relatively high clustering performance. One of the problems of the conventional clustering methods is poor clustering performance when the data distributions are overlapped and when the data distribution is concave shape, not convective. Also Genetic Algorithm: GA clustering with fitness function of Fisher distance between clusters or some other definition of distance is proposed.

The clustering method proposed here is GA based clustering with merge and split of the cluster. Therefore, it is possible to minimize fitness function (Fisher distance, in this case) through GA process. Also it is possible to refine the clusters created through GA process by merge and split process like ISODATA.

The following section describes the proposed GA based clustering with merge and split process followed by some experiments. Then conclusion is described together with some discussions.

II. PROPOSED METHOD

A. GA Based Clustering Method

GA allows minimization of fitness function which is defined as distance between clusters, in general. GA based clustering performance depends on the definition of fitness function.

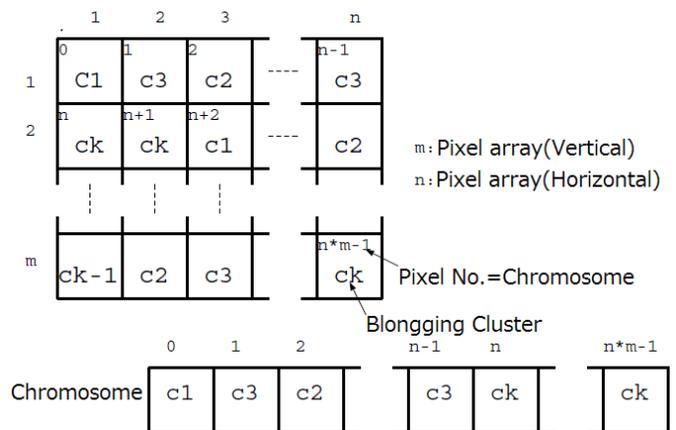Fundamental scheme for the GA based image clustering is shown in Figure 1.

Fig. 1. Fundamental scheme for the GA based clustering

Image data is expressed with two dimensional m by m matrixes. Chromosome is defined as pixel, at first. Then selection of chromosome has to be done (based on elite selection strategy in this case). Through cross over, and mutation processes, chromosome is updated by minimizing fitness function such as distance between clusters. Thus chromosome represents cluster number, finally. Conversion can be defined with difference of chromosome between the previous and the current chromosomes. The proposed clustering method introduces Fisher distance (Ratio of within cluster variance and between cluster variance) as fitness function.

B. Problem Discription

Basically, GA provides one of local minima, not global optimum solution. Therefore, clustering performance is not good enough. In particular, data distributions in feature space are overlapped as shown in Figure 2. It should consist of three clusters. It, however, they are overlapped each other. Therefore, it is not easy to make clusters. Furthermore, data distributions are concave shape, not convective shape Figure 3 shows conceptual illustration of process flow of the proposed merge and split procedure. In this case, cluster #1 is distributed as concave shape. If GA based cluster is applied to the data with the number of cluster is two. Then cluster #1 includes not only cluster #1 of data but also cluster #2 data at first. The number of clusters then increased from two to three. After that GA based clustering is applied to the data results in three clusters.

Finally, two clusters are merged to cluster #1 results in appropriate two clusters.
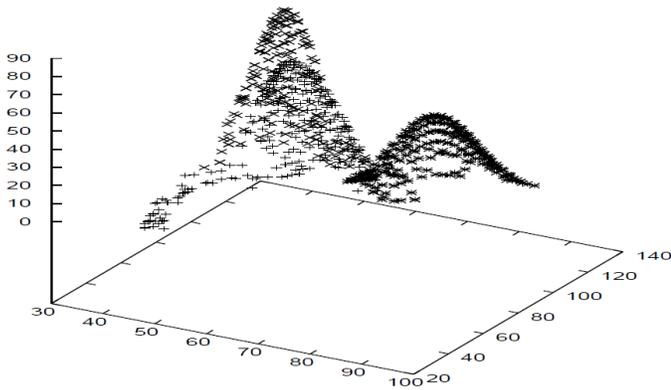

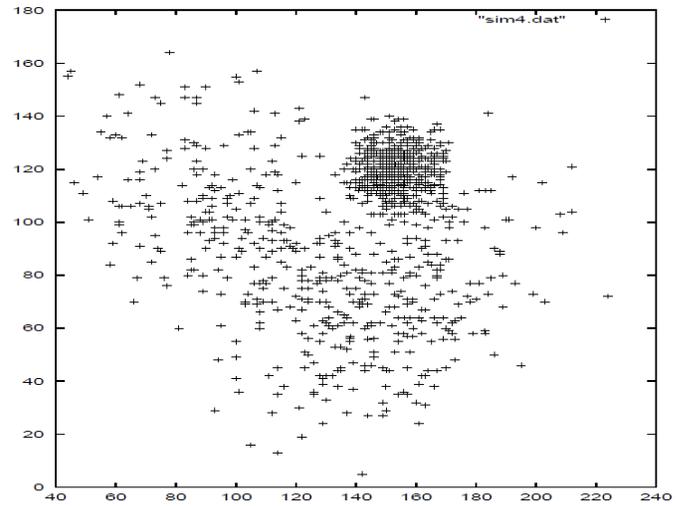Fig. 2.   Well overlapped data distribution in feature space

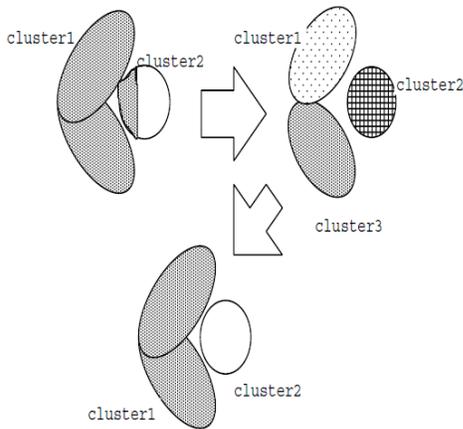Fig. 4.   Data distribution of two clusters in feature space

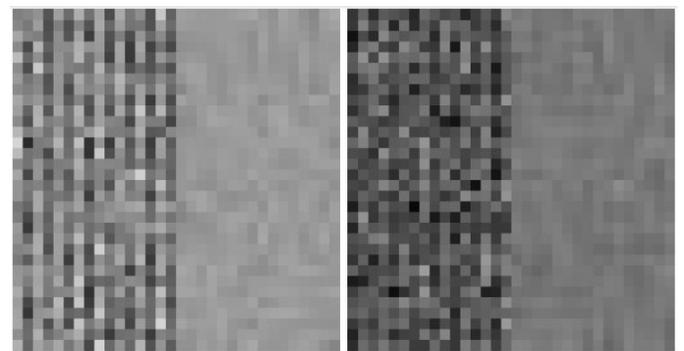Fig. 3.   Example of concave shaped data distribution in feature space.

(a)Band 1   (b)Band 2

Fig. 5.   Two clusters simulation imagery data

III.   EXPERIMENTS

A. *Preliminary Simulations*

In order to create the aforementioned situation of data distribution like Figure 3, two clusters of simulation data is created. The simulation data is generated manually and is distributed as shown in Figure 4 in two dimensional feature spaces which are corresponding to two bands, Band 1 and Band 2. The image consists of 32 by 32 pixels. Two clusters of simulation imagery data of Band 1 and Band 2 are shown in Figure 5 (a) and (b), respectively.

The clustered result by k-means clustering is shown in Figure 6 while the clustered result from the proposed method is shown in Figure 7. In the proposed method, the parameters for GA are as follows,

Decreasing factor for elite selection strategy: 0.75
Cross over probability: 0.6
Mutation probability: 0.03,
Range for iteration: 1000 to 3000

Meanwhile, clustered resultant images for k-means and the proposed GA based clustering methods are shown in Figure 8.
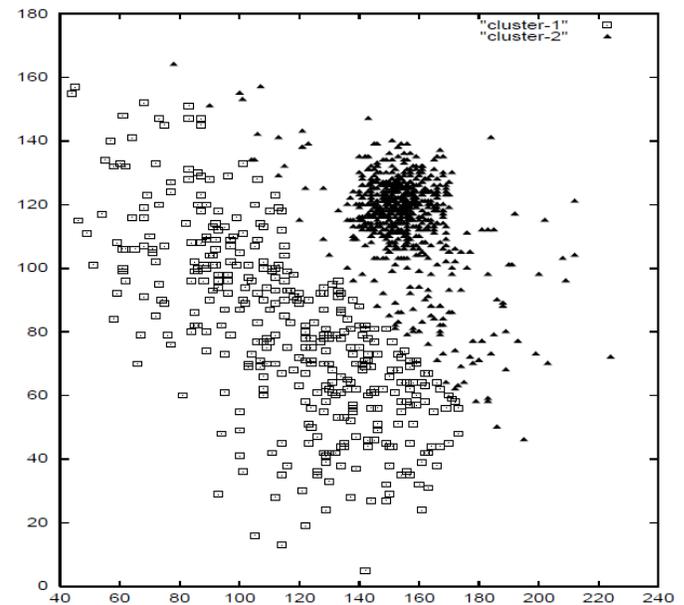

Fig. 6.   Clustered result by k-means clustering method on two dimensional feature plane (Percent Correct Clustering: PCC is 86.5%)

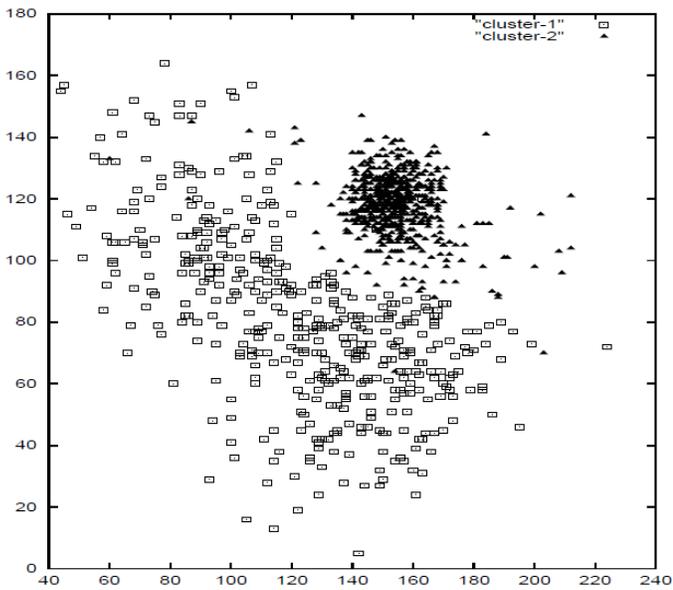Fig. 7. Clustered result by the proposed GA based method on two dimensional feature plane (PCC=93.1%)

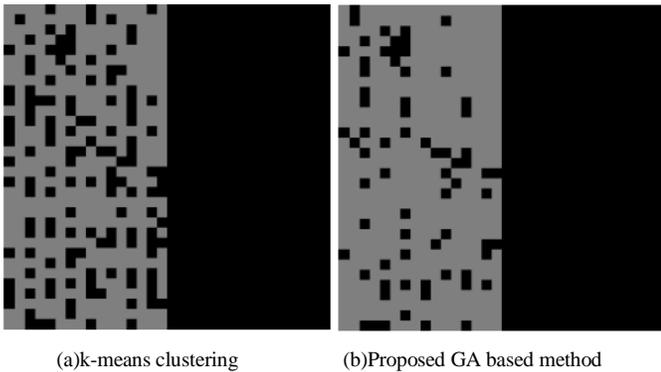(a)k-means clustering   (b)Proposed GA based method

Fig. 8. Clustered image data

Relation between Fisher distances after and before the merge and split process is shown in Figure 9. Fisher distance is increased remarkably by the GA algorithm.
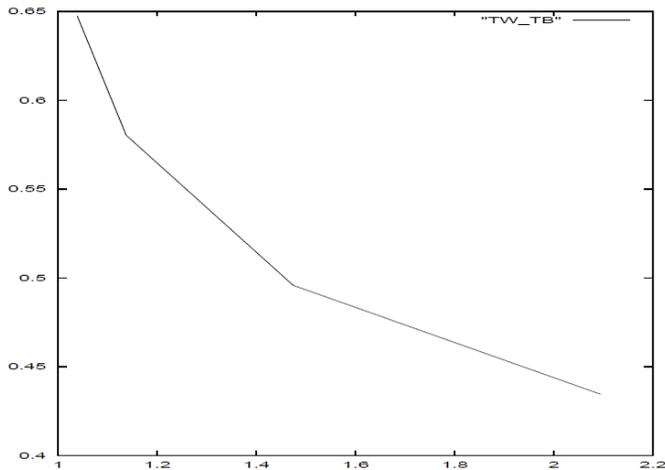

Fig. 9. Relation between Fisher distances after and before the merge and split process

B. Experiments with Remote Sensing Satellite Images

Landsat-5 TM image data of Saga city, Kyushu Japan which is acquired on 21 May 1985 is used for experiments. Figure 10 shows just two bands of images while Figure 11 shows data distribution on two dimensional feature plane. The number of clusters is set as two.
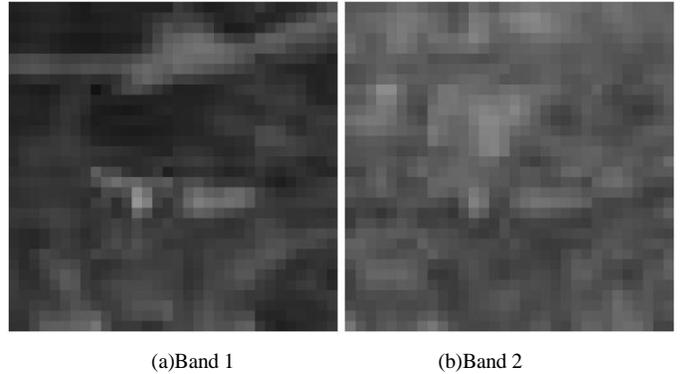

(a)Band 1   (b)Band 2

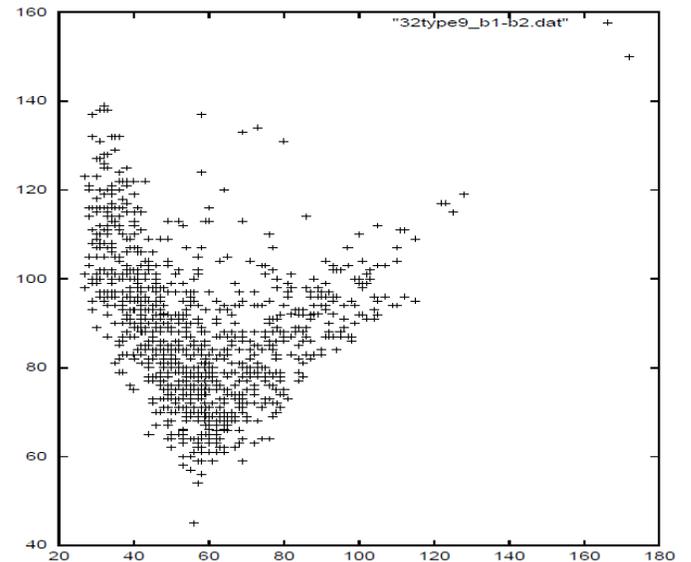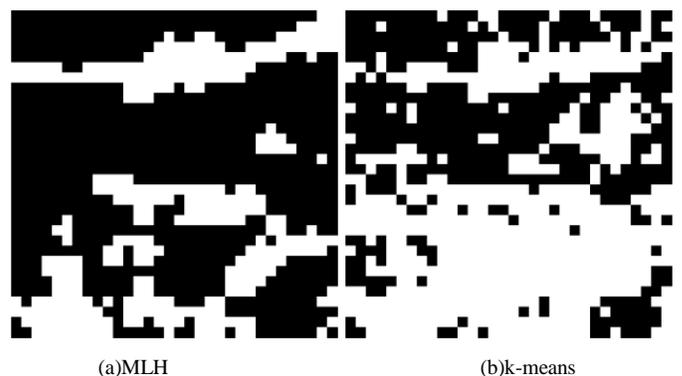Fig. 10. Landsat-5 TM image used for experiments (Original image #1)

Fig. 11. Data distribution on two dimensional feature plane.

Maximum Likelihood classification: MLH is applied to the data, first. Then k-means, ISODATA and the proposed GA based clustering methods are applied. Figure 12 shows the clustered results of these methods.

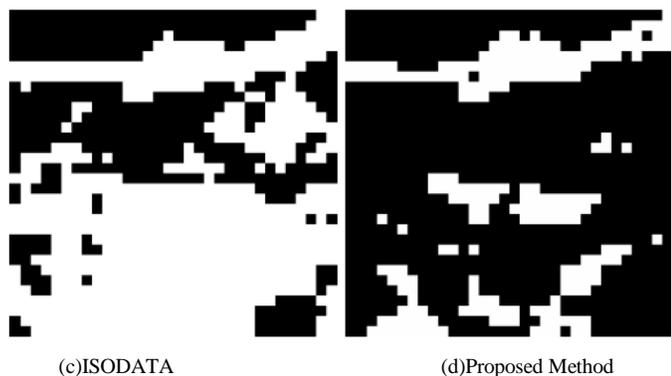(a)MLH   (b)k-means

(c)ISODATA   (d)Proposed Method

Fig. 12. Clustered images

The clustered resultant image of the proposed GA based method is very similar to that from MLH classification method while these from the conventional clustering methods, k-means and ISODATA differ from that from MLH classification method. On the other hand, the clustered result on the two dimensional feature plane for k-means, ISODATA, and the proposed methods are shown in Figure 13.

Same experiment is conducted with the different portion of the same satellite image data. Images of Band 1 and 2 are shown in Figure 14. Data distribution on two dimensional feature plane is shown in Figure 15. Using this image data, comparison among the aforementioned three clustering methods is conducted. k-means, ISODATA and the proposed GA based clustering methods are applied. Figure 16 shows the clustered results of these methods.

The clustered resultant image of the proposed GA based method is very similar to that from MLH classification method while these from the conventional clustering methods, k-means and ISODATA differ from that from MLH classification method. On the other hand, the clustered result on the two dimensional feature plane for k-means, ISODATA, and the proposed methods are shown in Figure 17.

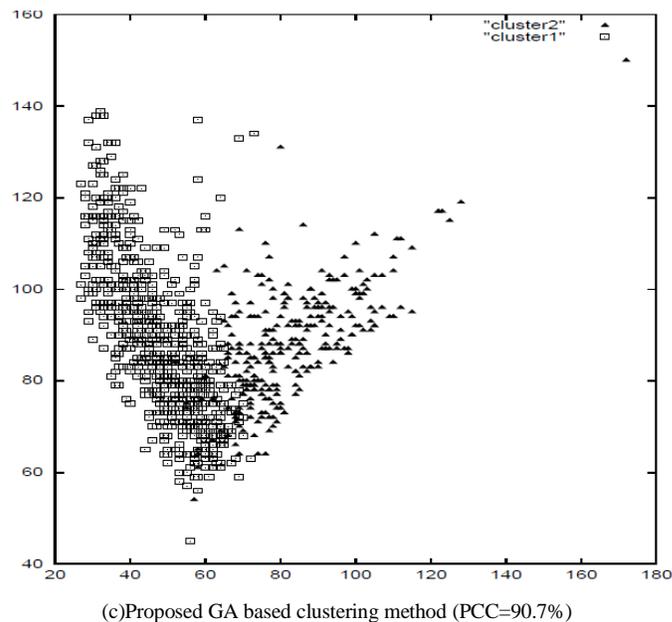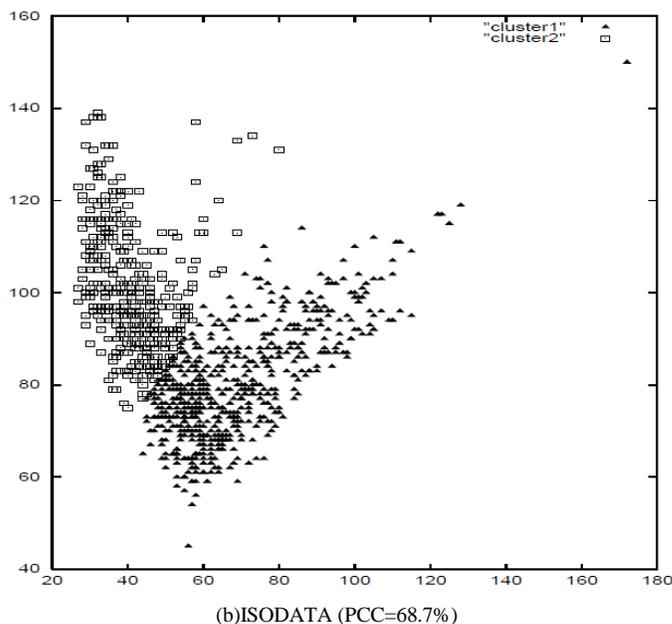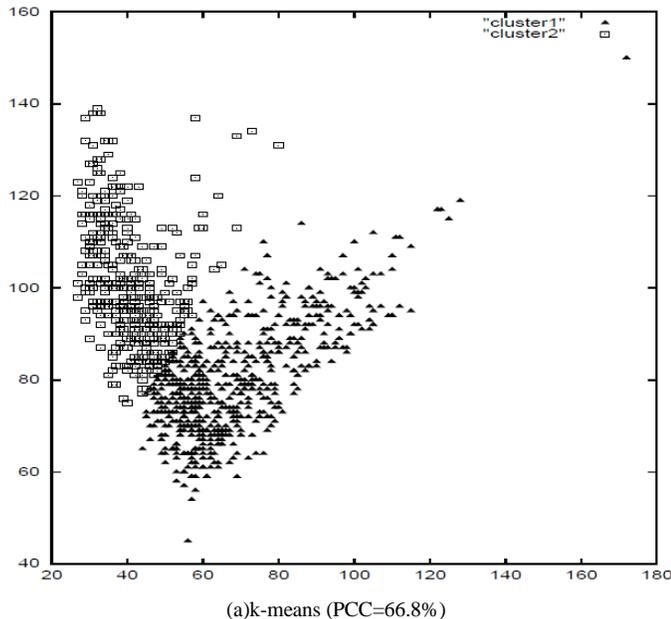
(b)ISODATA (PCC=68.7%)

(c)Proposed GA based clustering method (PCC=90.7%)
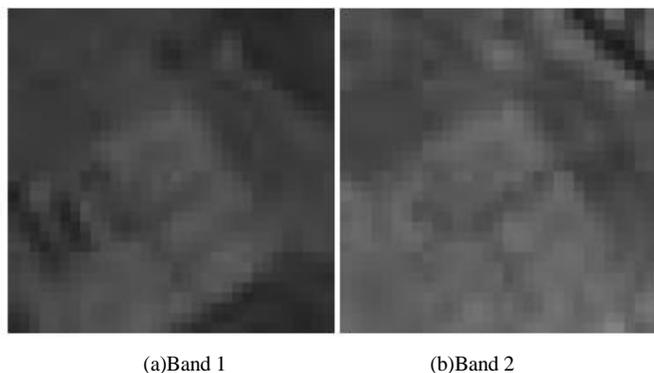
Fig. 13. Clustered results on two dimensional feature plane
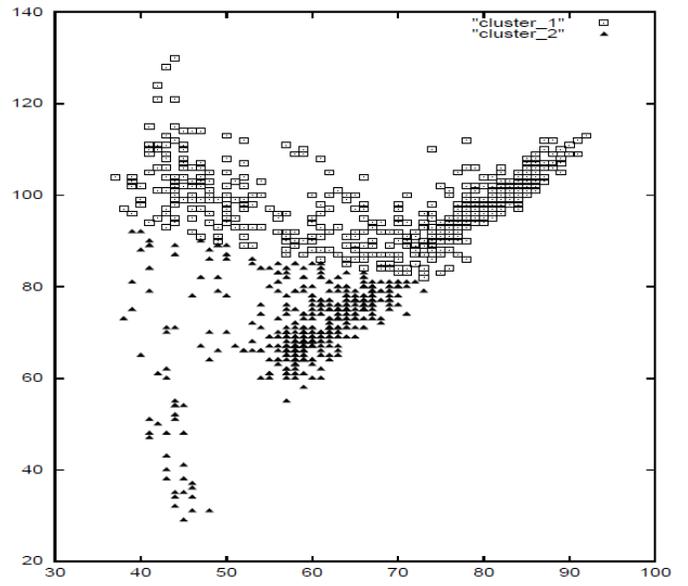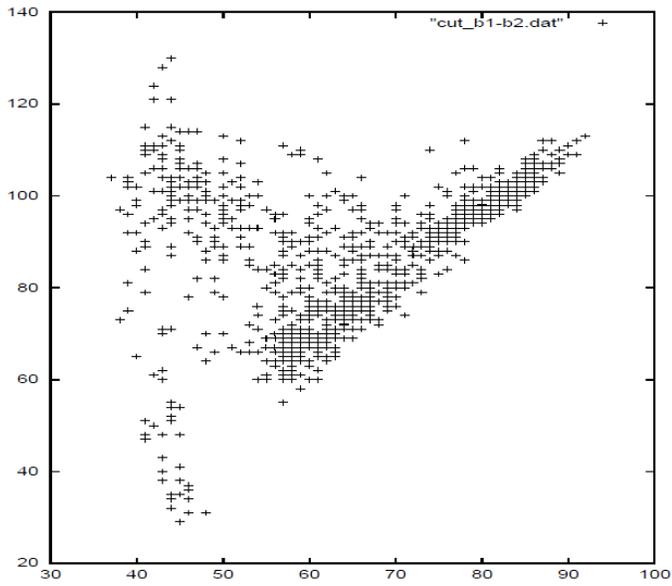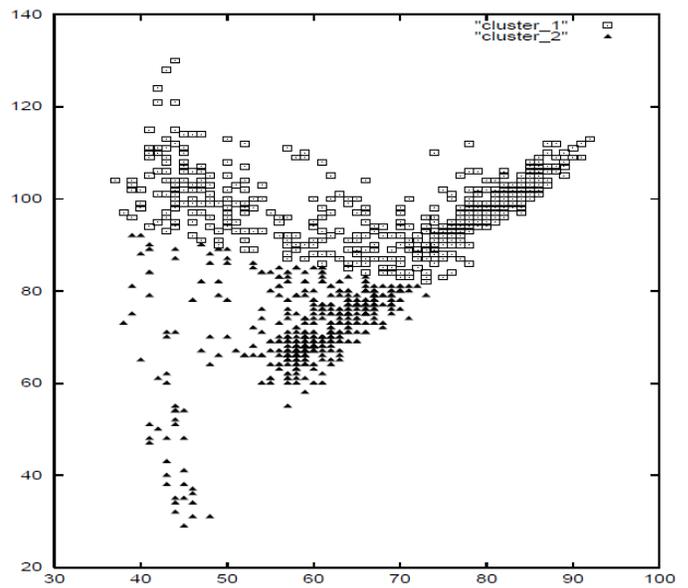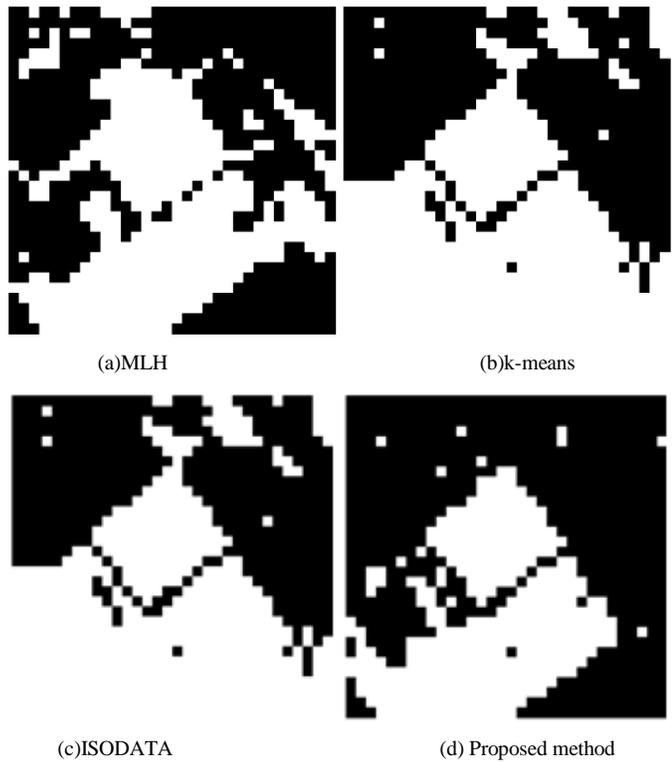

(a)k-means (PCC=66.8%)

(a)Band 1   (b)Band 2

Fig. 14. Another portion of Landsat-5 TM image data (Original image #2)

Fig. 15. Data distribution on two dimensional feature plane

(a)k-means (PCC=64.6%)

(a)MLH

(b)k-means

(b)ISODATA (PCC=64.7%)

(c)ISODATA

(d) Proposed method

Fig. 16. Clustered image
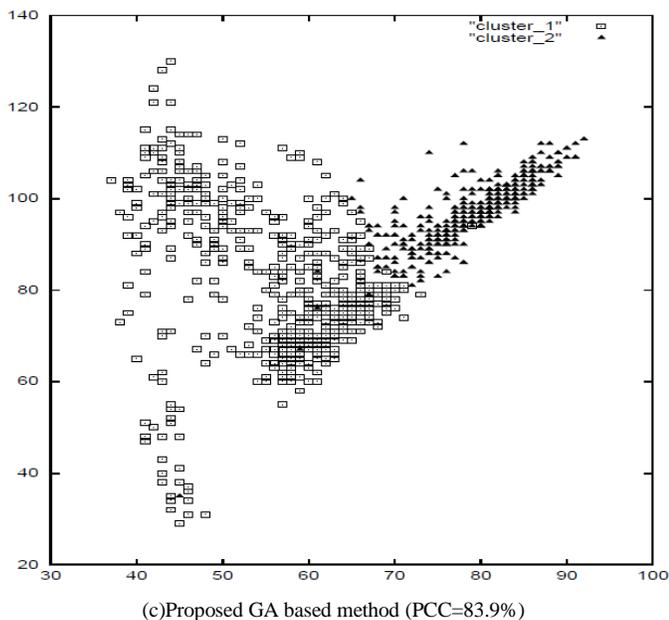
(c)Proposed GA based method (PCC=83.9%)

Fig. 17. Clustering results of data distribution on feature plane

For the proposed GA based method, between cluster variance depends on iteration number, obviously. Figure 18 (a) and (b) shows the between cluster variances for the original image #1 and #2, respectively.

C. Comparison with Geographical Map

Comparison between geographical map and the clustered resultant images is conducted. Figure 19 (a) shows the geographical map while Figure 19 (b) shows Landsat-5 TM image. Figure 20 shows the clustered images by the proposed method with and without merge and split processes. The clustered image by the proposed method with merge and split is much similar to the geographical map data than that by the proposed method without merge and split processes. Therefore, it is said that the proposed GA based clustering method is superior to the other conventional clustering methods, k-means and ISODATA as well as the proposed GA based clustering method without merge and split processes.

(a)Original image #1
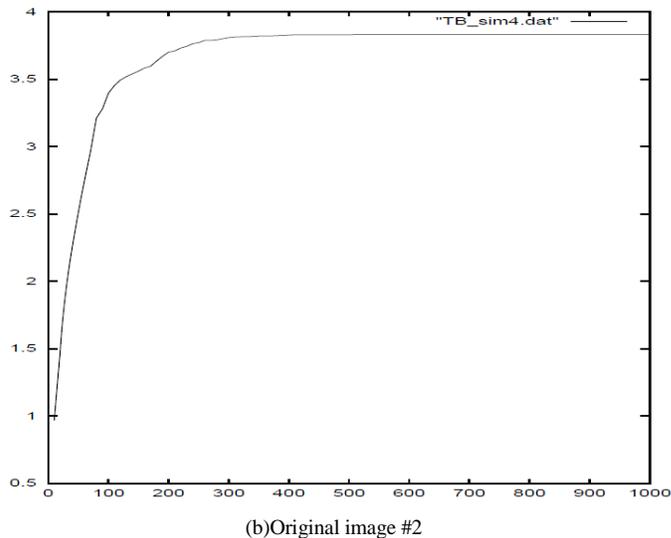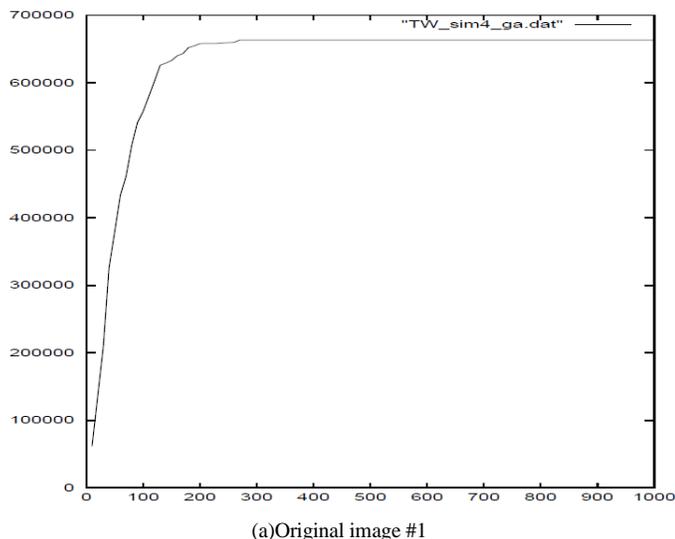
(b)Original image #2

Fig. 18. Between cluster variances for the original images 1 and 2

(a)Geographical map of Saga city, Japan

(b)Landsat-5 TM image of Saga city, Japan

Fig. 19. Comparison between Landsat-5 TM image and the corresponding geographical map
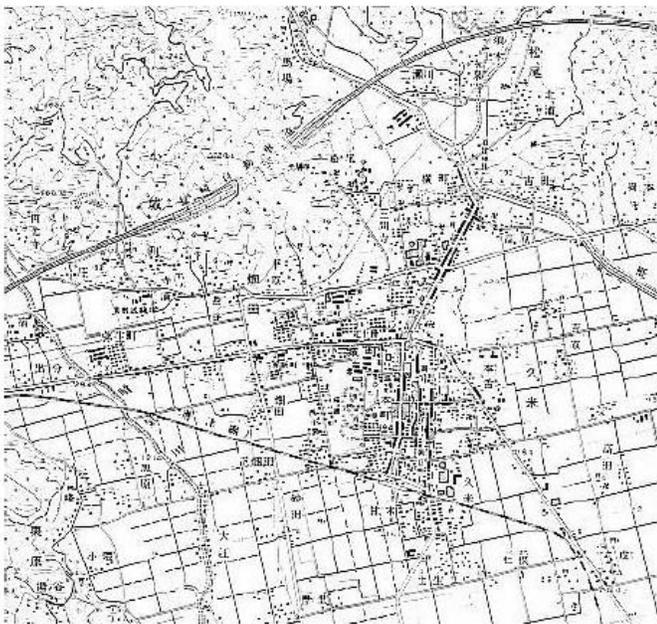
(a)Portion of geographical map

(b)Without merge and split   (c)With merge and split

Fig. 20. Clustered images by the proposed method with and without merge and split processes

IV. CONCLUSION

Genetic algorithm utilizing image clustering with merge and split processes which allows minimizing Fisher distance between clusters is proposed. Through experiments with simulation and real remote sensing satellite imagery data, it is found that the proposed clustering method is superior to the conventional k-means and ISODATA clustering methods in comparison to the geographic maps and classification results from Maximum Likelihood classification method.

The clustered image by the proposed method with merge and split is much similar to the geographical map data than that by the proposed method without merge and split processes. Therefore, it is said that the proposed GA based clustering method is superior to the other conventional clustering methods, k-means and ISODATA as well as the proposed GA based clustering method without merge and split processes.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

AUTHORS 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 332 journal papersAuthors Profile …