Remote Sensing Satellite Image Clustering by Means of Messy Genetic Algorithm

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

***Abstract*—Messy Genetic Algorithm (GA) is applied to the satellite image clustering. Messy GA allows to maintain a long schema, due to the fact that schema can be expressed with a variable length of codes, so that more suitable cluster can be found in comparison to the existing Simple GA clustering. The results with simulation data show that the proposed Messy GA based clustering shows four times better cluster separability in comparison to the Simple GA while the results with Landsat TM data of Saga show almost 65% better clustering performance.**

***Keywords*—Genetic Algorithm: GA; Messy GA; Simple GA; clustering introduction**

I. Introduction

As unsupervised image classification based on statistical methods, quantification theory and clustering are mentioned as typical methods [1]. Clustering methods can be broadly classified into hierarchical clustering, which treats pixels as individuals and classifies sets of individuals hierarchically, and non-hierarchical clustering, which divides a set of individuals at a time into a certain number of divisions [2]. The latter method is a method in which the initial cluster is given, the cluster to which the individual belongs is determined based on the distance between the cluster and the individual, the cluster centroid is obtained, and the individual is rearranged. The former differs from the latter in that clusters are formed based on the distance between individuals, between individuals and within and between clusters without giving an initial cluster [4]. In general, the latter method is frequently used because of faster convergence. In particular, when relocating, k clusters, i.e., k-means method [5] and ISODATA (Iterative Self Organizing Data Analysis Techniques A) [3] is famous.

In any clustering method, when n individuals are divided into k clusters, there is no guarantee that an optimal division result will be obtained. The latter can also be considered as a kind of optimal combination problem, and one of the effective methods is a Genetic Algorithm (GA) [6]. Clustering by GA, which effectively uses the effects of stochastic search and learning, is a method of improving the division by giving the evaluation criteria and initial division of the cluster. As a conventional method, there is clustering by Simple-GA [7].

However, since the location of the Simple-GA on the chromosome coincides with that on the chromosome, it is likely that the schema is superior or inferior depending on the location on the chromosome, and that the long schema is likely to be destroyed. On the other hand, Messy-GA is a variable-length list structure in which the chromosome-one gene expression is called a codon (locus allele). Therefore, it is unlikely that the schema will be superior or inferior due to the correspondence between the locus and the position on the chromosome, and the long schema can be expected to be preserved [7]. When applying the genetic algorithm to image clustering, cluster numbers are assigned by random numbers according to the pixel array, and the cluster number array (schema) effective for maximizing the fitness function is saved, crossed over, and mutation is performed. Probably searches for the optimal cluster while waking up, but originally the image has high spatial correlation, so the cluster to which the target pixel belongs is likely to match the cluster around it.

The author proposed a clustering method that takes such contextual information into account [8]. This paper further proposes a method using Messy-GA to guarantee schema preservation. The author takes up the degree of separation between clusters as the clustering accuracy, evaluate it using simulations and real satellite images (Landsat TM), and confirms the effectiveness of the proposed method.

II. Proposed Method

The author proposes clustering using Messy-GA in comparison with Simple-GA.

A. Messy GA Clustering

The schema length of the Simple GA is fixed. Therefore, relatively long schema which is effective for cross over is used to be broken. Consequently, it is difficult to find the most appropriate solution of chromosome. Meanwhile, cross over is much effective for Messy GA due to the fact that all the possible chromosome of maximum length can be prepared because the chromosome length is variable together with list of structural representation of chromosome. (1) Coding of chromosome, then initial pair of pixels number and cluster number is se.t (2) Fitness function evaluation. (3) Initialization. (4) Primordial phase. (5) Juxtaposition phase When the iteration number and the data number is exceed the threshold, all the pixels are assigned to cluster number.

B. Chromosome-Genotype Expression

Gene expression representing the state of dividing n individuals into k clusters is performed as shown in Table I.

Gene: $C_i = 0,1,\ldots, k-1$

That is, a pixel is defined as an individual, and its cluster number is defined as a gene. Genes are arranged according to the order of pixel arrangement, and GA is used in an algorithm

for stochastically searching for an optimal cluster of the pixel. At this time, the optimal cluster (remains) under the condition to maximize the fitness function shown as follows.

Gene sequence or a partial sequence of a certain part of the chromosome and another partial sequence (schema) at a certain probability, or by causing a mutation at a certain probability.

Find the best cluster. In the case of Simple-GA, since the schema description is fixed length, even if a valid schema for maximizing the fitness function can be searched, it is highly likely that it will be destroyed, but Messy-GA since the description of the schema is variable, the schema determined to be valid can be stored. This mechanism is shown in Fig. 1.

C. Fitness Function

The coding of the chromosome representing the division state in which n individuals are divided into *k* clusters is performed as follows:

$$((x_{i1}\ v_{i1})\ (x_{i2}\ v_{i2}) \ldots\ (x_{in}\ v_{in})) \quad (1)$$

where, assuming that the length of the chromosome in Simple-GA is 1,

$$1 > \ldots\ x_{i1},\ x_{i2},\ \bullet\ c,\ x_{in} > \ldots\ l \quad (2)$$

and allow a variable length. x_{i1} ... x_{in} indicates a locus, and v_{i1} ... v_{in} indicates an allele value at the locus.

A fitness function is defined by equation (3).

$$S_B\,(k) = S_T\text{-}S_W\,(k) \quad (3)$$

where S_T: sum of squares of *n* individuals,

$$S_r = \sum_{i=1}^{n} \|x_i - m\|^2 \quad (4)$$

S_W (*k*): sum of square sums in a cluster of *k* clusters,

$$S_w(k) = \sum_{i=1}^{n} \sum_{x_i \in C_i} \|x_i - m\|^2 \quad (5)$$

S_B (*k*): sum of inter-cluster sum of squares of *k* clusters,

$$S_B(k) = \sum_{i=1}^{n} \sum\ n_i \|m_i - m\|^2 \quad (6)$$

The *n* individuals are divided into *k* non-empty, mutually exclusive clusters.

D. Selection / Selection Operation

The same number of chromosomes as the population of the previous generation are selected from the population of the previous generation by using the expectation strategy using uniform random numbers and elite preservation strategy according to the fitness. At the same time, selection is performed[1]. As an example, the decrease in the expected value in the expected value strategy is 0.75.

TABLE. I. GENE EXPRESSION

Pixel_No.	0	1	…	n-1
Gene(Cluster_No.)	C_0	C_1	…	C_{n-1}

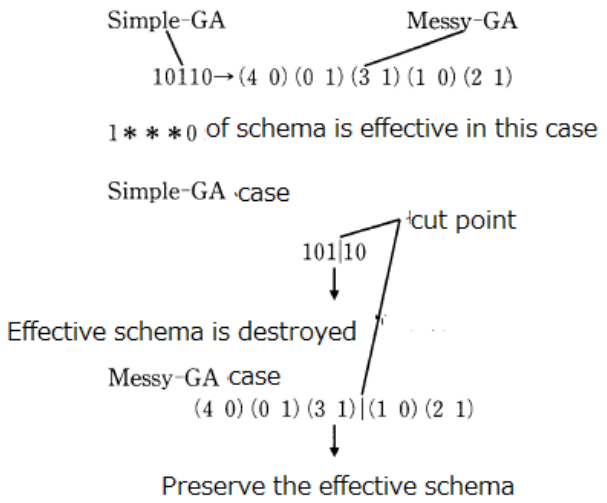


Fig. 1. Difference between Simple GA and Messy GA.

E. Crossover

The crossover operation performs multipoint crossover using dominant inheritance as a model. According to the crossover probability, cross-symmetric chromosomes are selected from the chromosomes selected in the selection and selection operation, and crossover is performed in the selected order.

At the time of crossover, the genotype mismatch between the two chromosomes occurs. Therefore, the reference locus is selected from the chromosomes selected as the crossover target using uniform random numbers. Based on the selected reference loci, the alleles are replaced according to equation (7), and the genotype matches.

$$j = i\text{-}h \pmod{n}, \qquad i = 0,1, \bullet\ c, n\text{-}1 \quad (7)$$

In this way, multipoint crossover is performed on chromosomes with unified allele types. However, actual allele replacement is performed only when the fitness of the replaced chromosome is improved. The chromosome where the allele replacement is performed updates the fitness every time the replacement is performed[2]. As an example, the crossover probability is 0.6.

F. Mutation Operation

According to the mutation probability, the chromosomal locus causing the mutation is randomly determined, and the allele is determined using the uniform random number at the determined chromosomal locus. If the fitness is improved when replacing with the determined allele, allele replacement is performed[3]. As an example, the mutation probability is 0.03.

G. Convergence Condition

The initial division of the cluster is set to 0 generation, and updating of the set generation is used as the program termination condition[4]. As an example, the end setting generation is 300,000 generations.

[1] As an example, the decrease in the expected value in the expected value strategy is 0.75.

[2] As an example, the crossover probability is 0.6.
[3] As an example, the mutation probability is 0.03.
[4] As an example, the end setting generation is 300,000 generations.

III. EXPERIMENT

The conventional method and the proposed method were applied to the simulation and actual satellite images, and the respective clustering accuracy was evaluated. Here, to show the superiority of the proposed method, the parameters of GA in the conventional method are the same as those of the proposed method.

A. Simulation Parameters

GA parameters are as follows.

- Early chromosome group 50
- Crossover probability 0.75
- Number of end generations 300,000
- Mutation probability 0.03 (Simple-GA only)

The simulation data creation parameters are as follows. Cluster individuals were generated by improving Neyman-Scott's method [9].

- 100 clusters · 2 bands · 2 clusters

Cluster average vector: $\mu_1 = (0.2, 0.9)^t$, $\mu_2 = (0.4, 0.6)^t$

- Cluster standard deviation $\sigma = 0.04$
- Distance between clusters 4σ

A population of 100 means an image of 10 × 10 pixels. 900 kinds of simulation image data were generated by changing the initial value of random numbers. Here, the distance between clusters i and j is shown in equation (8).

$$d_{i,j}^2 = (x_i - x_j)^t D_{i,j}^{-1}(x_i - x_j) + \frac{n}{2} ln|D_{ji}| \quad (8)$$

where $D_{i,j}$ is the covariance matrix of i, j of the cluster, $|D_{i,j}|$ is its determinant, and n is the number of dimensions.

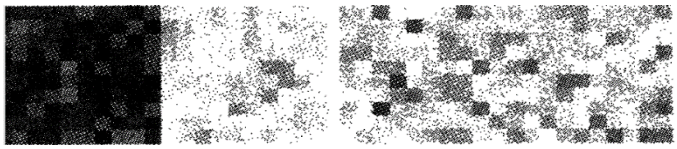
Normal random numbers were generated sequentially by giving the average and standard deviation, and constrained by the distance between clusters to construct a pixel array. Fig. 2 shows an example of the generated simulation image data. From the left, bands 1 and 2 of cluster 1 and bands 1 and 2 of cluster 2 are shown.

B. Simulation Results

Simulation images were generated to the extent that they could be considered as statistics (900 here). Clustering was performed by Simple-GA (SGA) and Messy-GA (MGA), and the degree of separation between clusters represented by equation (8) was evaluated. The cluster result images of Simple GA and Messy GA clustering are shown in Fig. 3, 4, respectively. Here the author shows only two of the 900 trials. At this time, Table II shows the number of generations up to convergence and the final degree of separation between clusters.

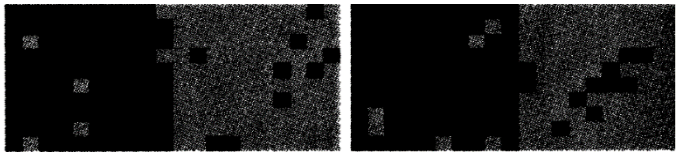
Fig. 5 shows an example of a change in the degree of separation between clusters (learning process).

Here, TBfitness and generation indicate the degree of separation between clusters and the number of convergent generations, respectively. In the figure, the broken line represents the learning process of SGA, and the solid line represents the learning process of MGA. All genotypes with deceptive order-length building blocks at the initialization stage because the chromosomes characteristic of MGA are of variable length and the chromosomes are composed of a list of loci and allele pairs can be generated and only genotypes with valid building blocks can be left to posterity. In addition, gene lists can be exchanged on the chromosome, and this optimization learning takes time. Chromosomes are significantly different from SGAs, which are represented by a fixed length.



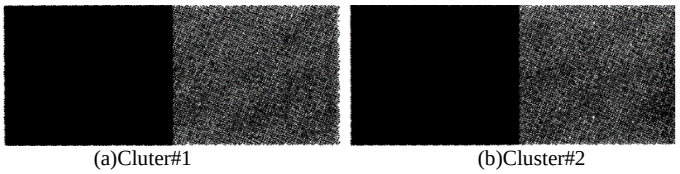
(a)Cluster#1Band 1 (b)Band 2 (c)Cluster#2Band 1 (d)Band 2

Fig. 2. Simulation Data used.



(a)Cluter#1 (b)Cluster#2

Fig. 3. Simple GA.



(a)Cluter#1 (b)Cluster#2

Fig. 4. Messy GA.

TABLE. II. THE NUMBER OF GENERATIONS UP TO CONVERGENCE AND THE FINAL DEGREE OF SEPARATION BETWEEN CLUSTERS

Method	Separability_Between_Clusters	Iteration_No.
SGA	779.5	471
MGA	2866.1	6947

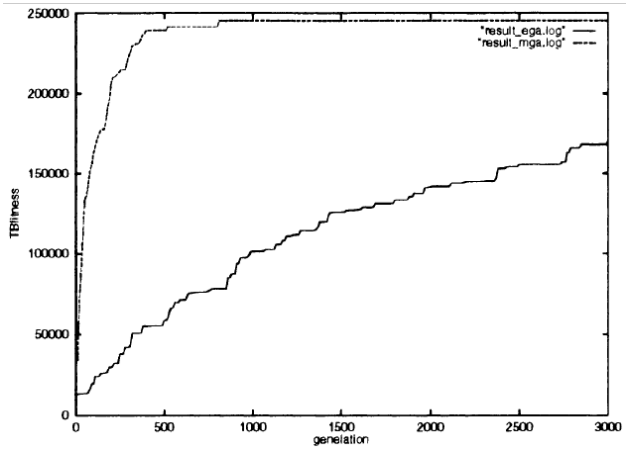


Fig. 5. An Example of a Change in the Degree of Separation between Clusters (Learning Process).

C. Satellite Image Results

Next, the results applied to real satellite images are shown. The GA operator parameters are as follows:

- Early chromosome group 50
- Crossover probability 0.75
- Number of end generations 300,000
- Mutation probability 0.03 (Simple-GA only)

The Landsat TM (Thematic Mapper) image around Saga city in March 25 1986 was used as an actual satellite image. Fig. 6 shows the location of intensive study area (Red circle) in the Google map.

From the image, a portion of 32 x 32 pixels in height and width is extracted as shown in Fig. 7. Landsat TM has the following seven spectral bands, including a thermal band:

Band 1 Visible (0.45 - 0.52 μm) 30 m

Band 2 Visible (0.52 - 0.60 μm) 30 m

Band 3 Visible (0.63 - 0.69 μm) 30 m

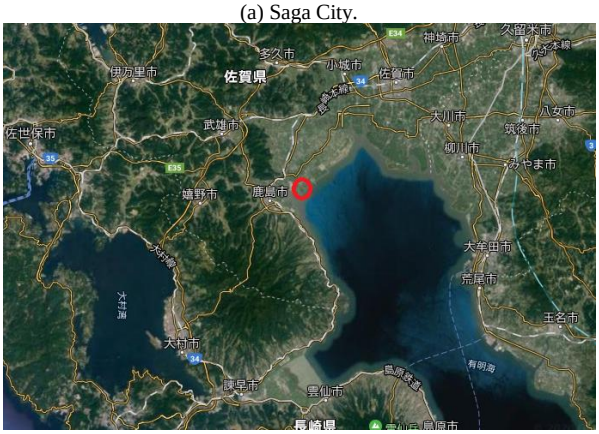
Band 4 Near-Infrared (0.76 - 0.90 μm) 30 m

Band 5 Near-Infrared (1.55 - 1.75 μm) 30 m

Band 6 Thermal (10.40 - 12.50 μm) 120 m

Band 7 Mid-Infrared (2.08 - 2.35 μm) 30 m

(a) Saga City.



(b) Intensive Study Area.

Fig. 6. Location of Intensive Study Area.

Fig. 7. Extracted Portion of Landsat TM Image.

Ground Sampling Interval (pixel size): 30 m reflective, 120 m thermal. 6 band data (excluding thermal band) are used for clustering. 5 clusters (urban, road, soi1, water, paddy) are assumed. Therefore, the number of clusters are set at five.

True color image of a portion of Landsat TM image which is acquired on March 25 1986 is shown in Fig. 8.

Fig. 9(a) shows only band 1 among the images used at that time. The resulting images by SGA and MGA clustering are shown in Fig. 9(b) and (c), respectively.

The clustered image of SGA shows noisy while that of MGA shows relatively smooth. In particular, nevertheless Ariake Sea area has to be clustered as one cluster, SGA result shows not only water body but also base soil, rice paddy, etc. Meanwhile, MGA result shows comparatively reasonable cluster.

Table III shows the number of convergent generations and the degree of separation between clusters.

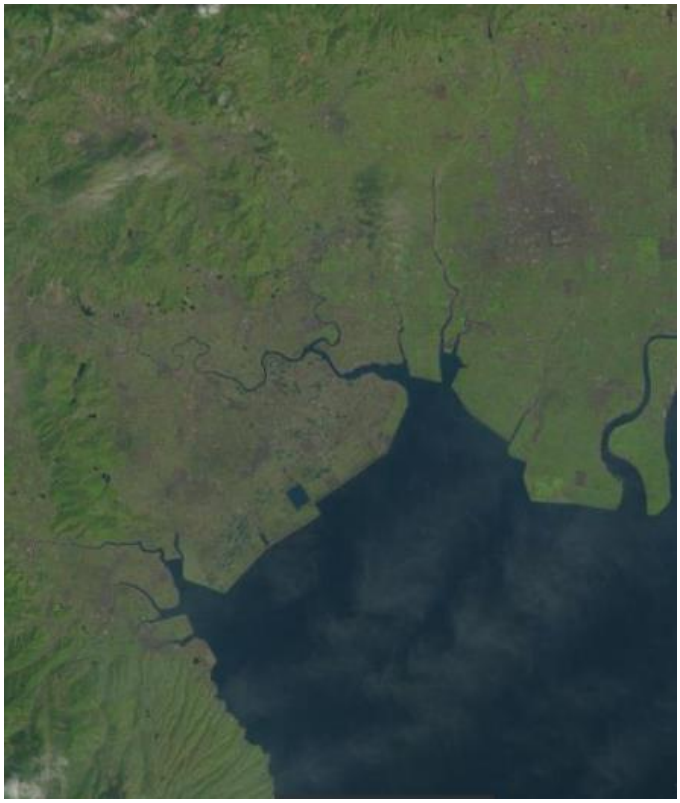
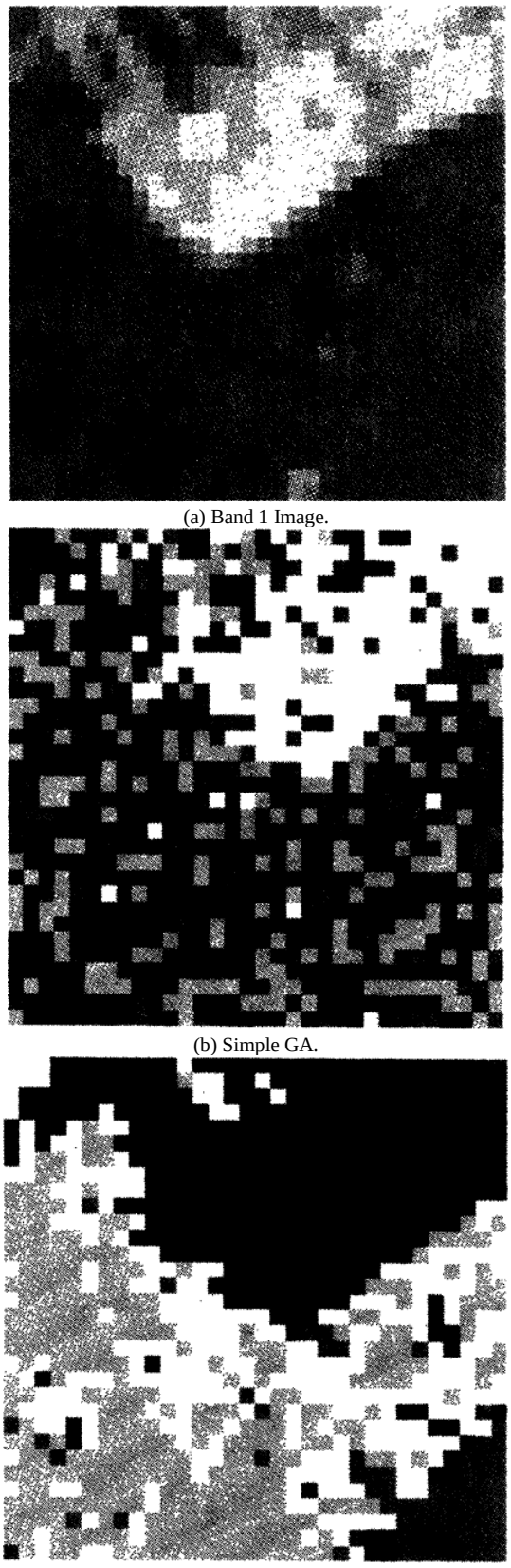


Fig. 8. A Portion of Landsat TM Image which is Acquired on March 25 1986.



(a) Band 1 Image.

(b) Simple GA.

(c) Messy GA.

Fig. 9. Band 1 of Landsat TM Image and the Resultant Images of Simple GA and Messy GA Clustering.

TABLE. III. CLUSTERING RESULTS USING LANDSAT TM IMAGES

Method	Separability_Between_Clusters	Iteration_No.
SGA	335	29238
MGA	554	299840

From these experimental results, it is found that Messy GA is superior to the conventional Simple GA from the view-points of reasonable clustered result and separability between clusters, the required time for clustering processes is much longer than Simple GA.

IV. CONCLUSION

Messy Genetic Algorithm (GA) is applied to the satellite image clustering. Messy GA allows to maintain a long schema, due to the fact that schema can be expressed with a variable length of codes, so that more suitable cluster can be found in comparison to the existing Simple GA clustering.

In Simple-GA, a gene has a fixed-length list structure, so a long schema is likely to be destroyed. In contrast, Messy-GA has a variable-length list structure and can store a long schema. As a result of comparing and evaluating the accuracy of the two clusters using 900 types of simulation data, the separation between clusters was shown to be about four times, and the result using Landsat TM image showed about 65% improvement, indicating that Messy-GA clustering turned out to be superior to Simple-GA. However, it was also found that the number of convergent generations was about 10 times higher for Messy-GA than for Simple-GA.

The author confirmed that both Simple-GA and Messy-GA surpassed the accuracy of k-means clustering, and also confirmed the tendency of accuracy improvement due to the difficulty of clustering, but the author will report these opportunities again.

V. FUTURE RESEARCH WORKS

Further experiments are required for validation of Messy-GA clustering effectiveness with the other remote sensing satellite images. Also, the applicability of the proposed Messy-GA clustering has to be attempted for not only remote sensing satellite image, but also the other images.

ACKNOWLEDGMENT

The author would like to thank former students of Dr. Akira Yoshizawa and Mr. Koichi Tateno as well as Professor Dr. Hiroshi Okumura of Saga University for his valuable comments and suggestions.

REFERENCES

[1] Mikio Takagi and Yuhisa Shimoda, edited by Kohei Arai, “Image Analysis Handbook”, University of Tokyo Press, 1991.

[2] Anderberg, M.R., Nishida (translation): Cluster analysis and its application, Uchida Ritsuruho, 1988.

[3] Ball, G.H. and D.J.Hall: ISO-DATA-Novel method of data: Analysis and patter classification, Menlo Park, California, Stanford Research Institute, 1965.

[4] Rance, G.N. and W.T.Williams: A general theory of classification sorting strategies-Hierarchical System-, Cognitive Journal, 9, 4, 373-380, 1967.

[5] Selim, S.Z. and M.A.Ismail: K-mean type algorithms: A general convergence theorem and characterization of local optimality, IEEE Trans.on PAMI-6, 1, 81-87, 1984.

[6] T. Kato and K. Ozawa: Non-hierarchical clustering using genetic algorithm, IPSJ Journal, 37, 11, 1950-1959, 1996.

[7] Masashi Iba: Basics of Genetic Algorithms, Ohmsha, 1994.

[8] Satoshi Yoshizawa, Kohei Arai: Clustering using genetic algorithm combined with spectrum and context information, Journal of the Institute of Image Electronics Engineers of Japan, 31, 2, 202-209, 2002.

[9] J. Neyman and E.L.Scott: Statistical approach to problems of cosmology, Journal of the Royal Statistical Society Series B 20,1-43, 1958.

AUTHOR'S 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 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 Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.htm