Clustering Method Based on Messy Genetic Algorithm: GA for Remote Sensing Satellite Image Classifications

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

Abstract— **Clustering method for remote sensing satellite image classification based on Messy Genetic Algorithm: GA is proposed. Through simulation study and experiments with real remote sensing satellite images, the proposed method is validated in comparison to the conventional simple GA. It is also found that the proposed clustering method is useful for preprocessing of the classifications.**

Keywords- clustering; classification; Genetic Algorityhm: GA; Messy GA; simple GA.

I. INTRODUCTION

Although there are conventional clustering methods such as k-means, ISODATA, etc., these cannot ensure the optimum clustering results at all [1]. Also Genetic Algorithm: GA based clustering method is proposed as an optimum combination problem solving method. It still cannot ensure the optimum clustering result [2]. GA based clustering uses probabilistic searching method for optimum combination through learning processes. The conventional GA based clustering is referred to “Simple GA”. Simple GA has problem such as relatively long schema used to be broken, fitness of schema may change in the defined chromosome in GA algorithm. Therefore, simple GA cannot ensure optimum clustering result at all.

In order to overcome the problems of simple GA based clustering method, Messy GA based clustering method is proposed in the paper. Messy GA utilize “Codon” which is defined as variable length of list structure of chromosome representation [3]. Therefore, relatively long schema used to be maintained and fitness of schema may not been changed in the defined chromosome in GA algorithm. Convergence processes are discussed [4]. Also Modified ISODATA clustering is proposed for acceleration of convergence processes [5]. GA based ISODATA is also proposed for improvement of clustering performance [6]. On the other hands, online clustering utilizing learning automata and pursuit reinforcement competitive learning is proposed [7],[8].

The following section describes the proposed Messy GA based clustering method followed by simulation study and some experimental study with remote sensing satellite images. Then conclusion is described together with some discussions.

II. PROPOSED METHOD

A. GA Clustering

Genetic Algorithm: GA based clustering is defined as follows,

(1) The image in concern is defined as two dimensional array of pixels which is shown in Figure 1.

(2) Chromosome is defined as a pair of pixel number and cluster number.

(3) Fitness function is defined as between cluster variance

(4) Cross over, mutation, and selection of the chromosome is repeated until the finish condition is satisfied though the processes that higher fitness function of chromosome is remained (This is the typical GA processes)

(5) Thus all the pixels are assigned the most appropriate cluster number (cluster results)

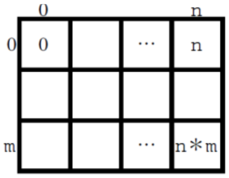
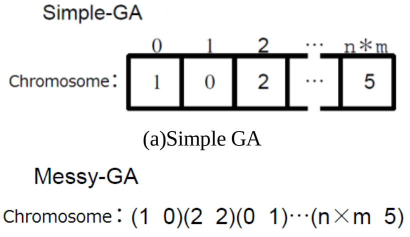


Figure 1 Definition of image (two dimensional array of pixels) in concern for clustering

In this process, chromosome is illustrated in Figure 2 (a) which is referred to Simple GA while Messy GA chromosome is shown in Figure 2 (b).



(a)Simple GA

(b)Messy GA

Figure 2 Chromosome definitions for Simple and Messy GA

Messy GA is composed with the following three phases,

(1) Initial phase:

All the schema of which chromosome with optimum scheme is generated referencing to the maximum schema length

(2) Primordial phase:

Selection of chromosome until the cross over process can be applied to chromosome through the following processes, 1) the maximum length of schema, k is determined, 2) all the possible schema is generated (The number of possible combination of schema is ${}_kC_m2^m$), 3) through fitness function of chromosome is evaluated, then 4) selection is made by tournament selection method

(3) Juxstapositional phase:

Choose the arbitrary pair of chromosome for cut off (Cut) and connect (Splice). The probability of the cut depends on the chromosome length. Meanwhile, splice is made independently the chromosome length. Therefore, splice is used to be made for the shorter chromosome because the probability of splice for shorter chromosome is less than that for long chromosome.

Figure 3 shows the process flow of the Messy GA.

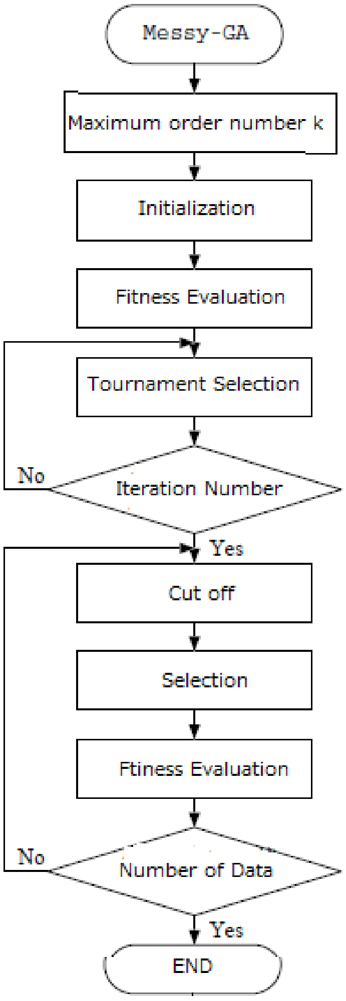


Figure 3 Process flow of Messy GA

B. Simple GA Clustering

The process of simple GA clustering is as follows,

(1) Pair of pixel number (ranges from 0 to n-1) and cluster number (ranges from 0 to k-1)is set as an initial stage

(2) Fitness function is evaluated with the between cluster variance

(3) Expectation strategy with uniformly distributed random number is used together with elite preserve strategy for selection of chromosome which has high score of the fitness function

(4) Multi point cross over, then applied to the previously preserved chromosome and the current chromosome with cross over probability

(5) The cross over processed chromosome is preserved only if the fitness function is greater than the previous before cross over is applied

(6) Mutation is applied with uniformly distributed random number with mutation probability

(7) Finish condition is set at the number of GA process is exceeded at 3000.

C. 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 structural representation of chromosome.

(1) Coding of chromosome, then initial pair of pixels number and cluster number is set

(2) Fitness function evaluation

(3) Initialization

(4) Primordial phase

(5) Juxstapositional phase

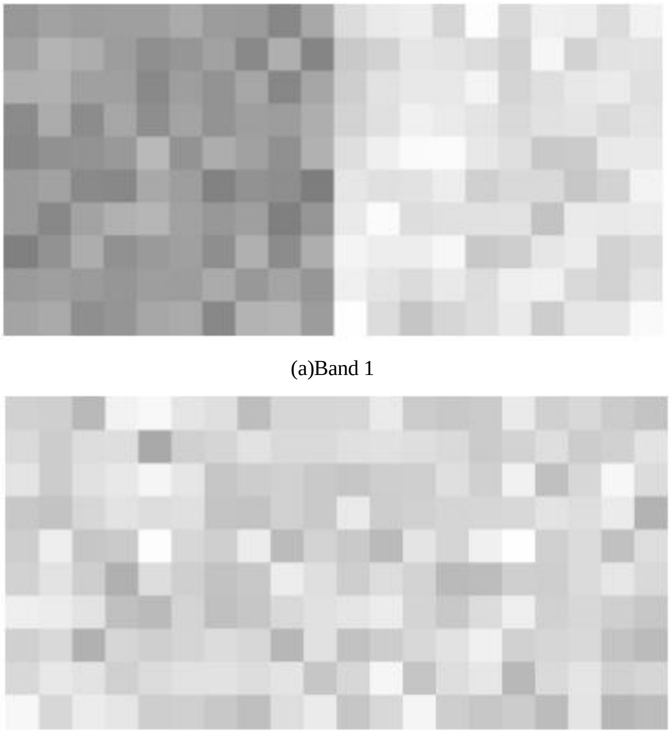
When the iteration number and the data number is exceed the threshold, all the pixels are assigned to cluster number.

III. Experiments

A. Simulation Results

Simulation study is conducted with the following parameters, The Initial chromosome assignment number=50, Cross over probability=0.75, Finalized generation number (iteration number)=3000, Mutation probability=0.03 (this is only for Simple GA), Standard deviation of cluster=0.04, Between cluster variance (distance between clusters)=0.16, The number of data=100, The number of bands of the data=2, The number of cluster=2.

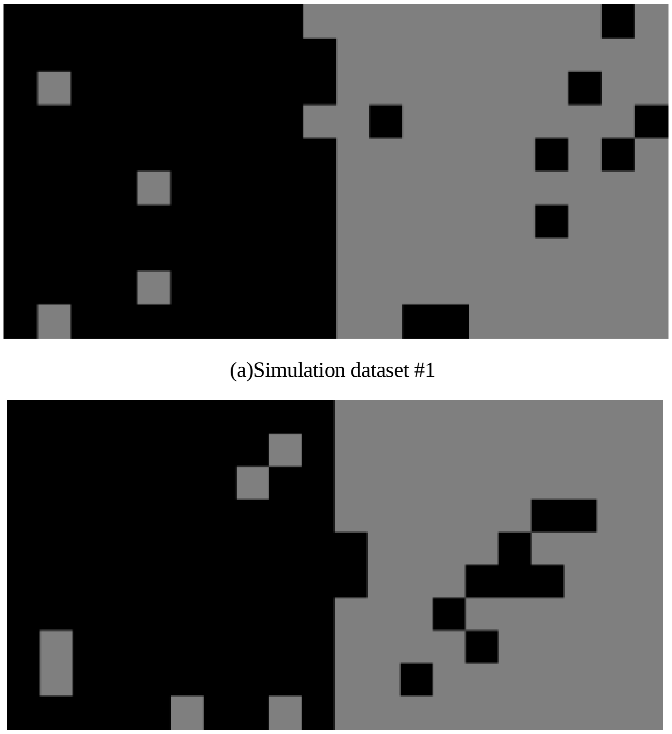
Figure 4 shows generated simulation dataset of band 1 and 2. There are two clusters, cluster # 1 (left) and # 2 (right). Figure 5 shows the clustering results from the Simple GA clustering while Figure 6 shows the clustering results from the Messy GA clustering.



(a)Band 1

(b)Band 2

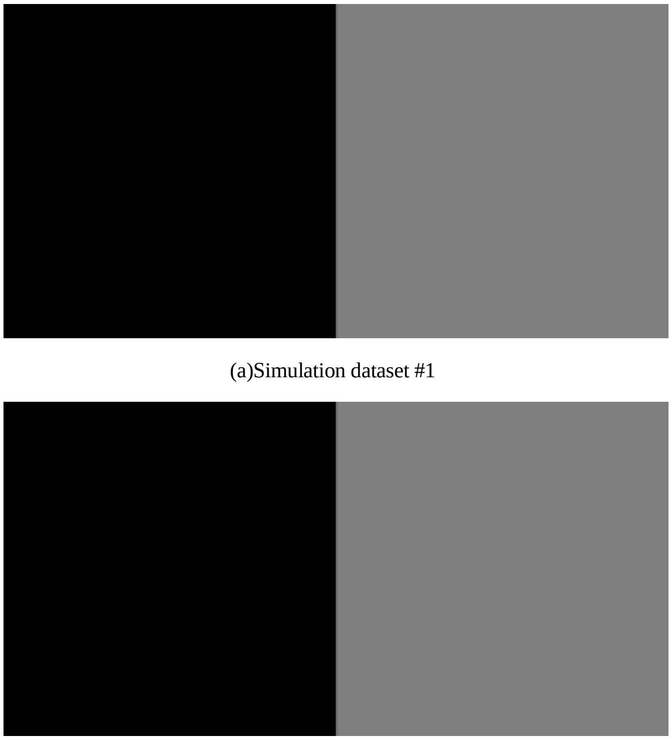
Figure 4 Simulation dataset used



(a)Simulation dataset #1

(b)Simulation dataset #2

Figure 5 Clustering results from the Simple GA clustering



(a)Simulation dataset #1

(b)Simulation dataset #2

Figure 6 Clustering results from Messy GA clustering

Such this simulation is repeated for 900 times. Then the clustering performance is evaluated. The convergence iteration number, between cluster variance, and percent correct clustering is evaluated as cluster performance.

These are shown in Table 1.

TABLE I. CLUSTERING PERFORMANCE OF THE SIMULATION STUDY

	Converged Iteration Number	Between Cluster Variance	Percent Correct Cluster
Simple GA	2866	171330	0.89
Messy GA	780	230947	0.99

All the evaluation items for Messy GA clustering is superior to those for Simple GA clustering. Significance test is also conducted for the clustering performance with 5% of significance level.

With the number of samples of 900, confidence interval at the confidence level of 95% is evaluated. The results show that the clustering results in Table 1 is significant.

Between cluster variance of fitness function is also evaluated as a function of iteration number. Figure 7 shows an example of the results from the between cluster variance evaluation.

Messy GA clustering (Dotted line in the Figure 7) converged faster than Simple GA clustering (Straight line in the Figure 7).

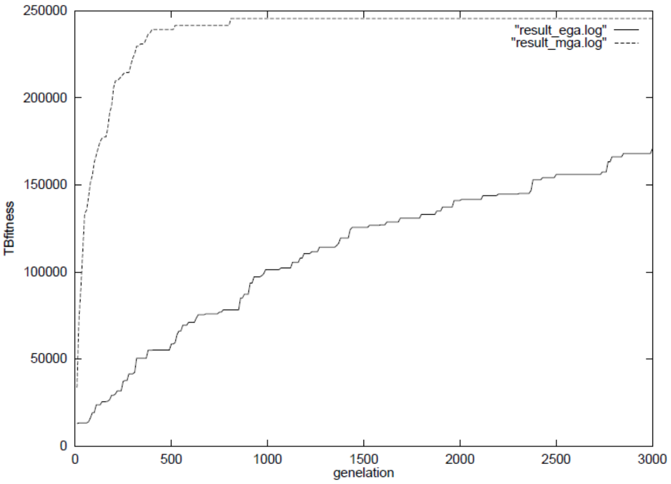
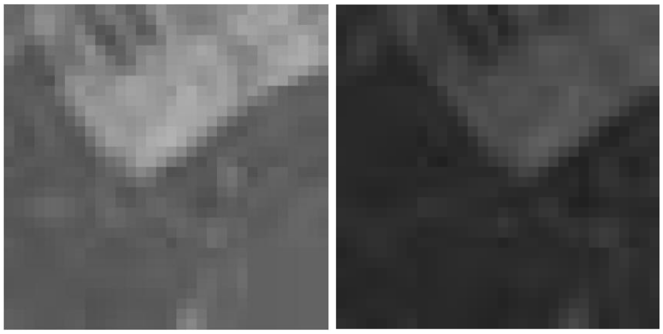


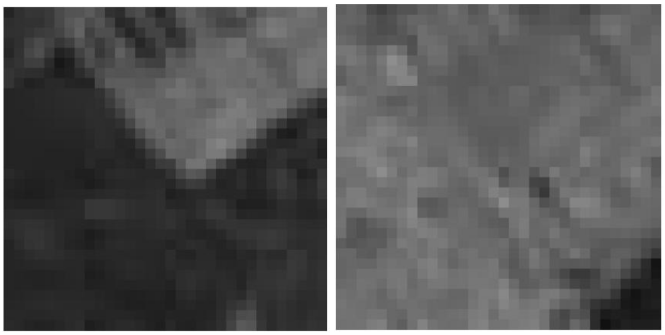
Figure 7 Example of the results from the between cluster variance evaluation.

B. *Experiemntal study*

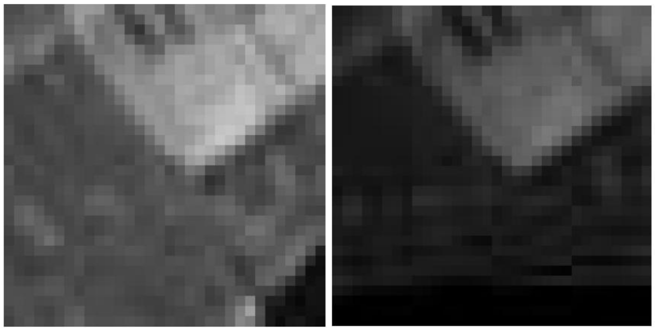
Small portion of LANDSAT-5/TM data (32 x 32 pixels) of the south-west of Saga city which is situated in the northern Kyushu, Japan which is acquired on May 28 1986 is used for the experiments. TM data consists of 6 bands, blue, green, red, near-infrared and two shortwave infrared bands with 30 m of Instantaneous Field of View: IFOV. Other than these, there is thermal infrared band with 120 m of IFOV. The thermal infrared band data is not used for the experiments. Figure 8 shows the images for the aforementioned 6 bands.



(a)Band 1 (b)Band 2



(c)Band 3 (d)Band 4



(e)Band 5 (f)Band 7

Figure 8 LANDSAT-5/TM Images of Saga city, Kyushu, Japan acquired on May 28 1986

GA parameters are set as follows,

The number of the initial chromosome=50,

Cross over probability=0.75,

Finished generation number=30000,

Mutation probability (only for Simple GA)=0.03,

The number of final clusters=5.

In order to evaluate clustering performance, maximum likelihood classification method is applied to the data with the following five classes, artificial structure, road, water body, paddy field, and bare soil. Classified results are shown in Figure 9. Also clustered results for Simple GA and Messy GA are shown in Figure 10 (a) and (b), respectively. The results are quite obvious that the clustered result from Messy GA is superior to that of Simple GA.

Clustering performance of Percent Correct Clustering: PCC and the number of correct clustering pixels are shown in Table 2 together with the converged iteration number.

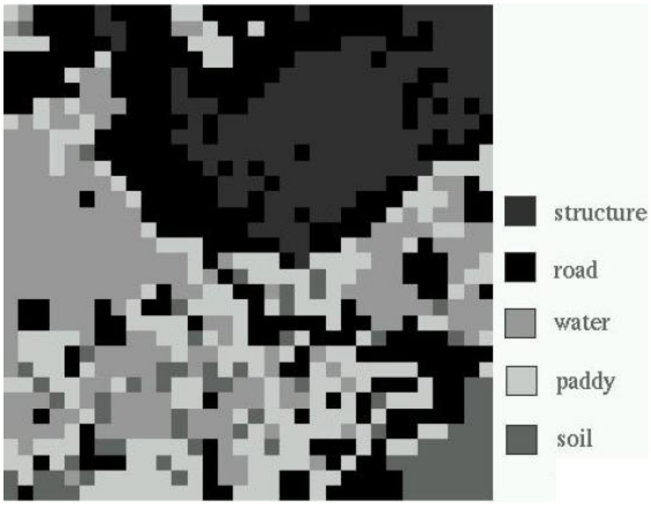
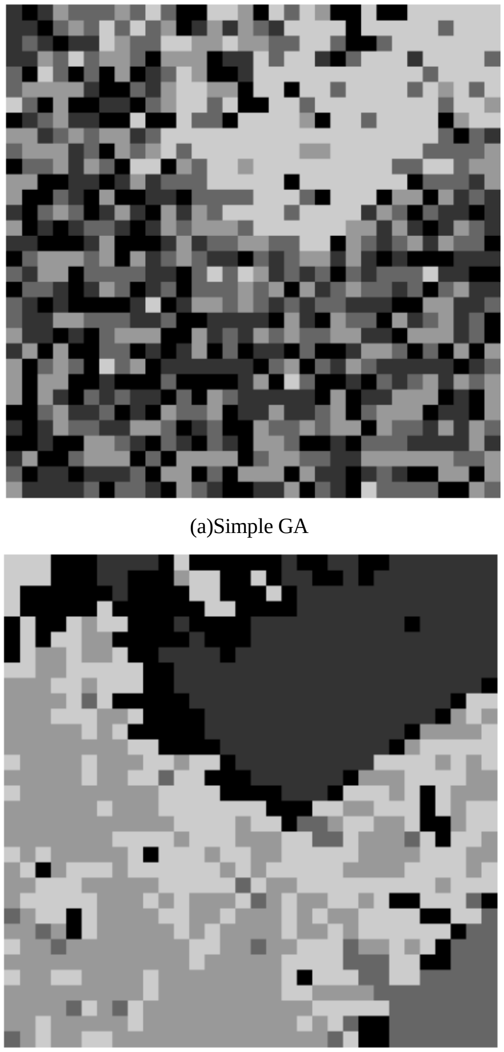


Figure 9 Classified results



(a)Simple GA

(b)Messy GA

Figure 10 Clustered Results

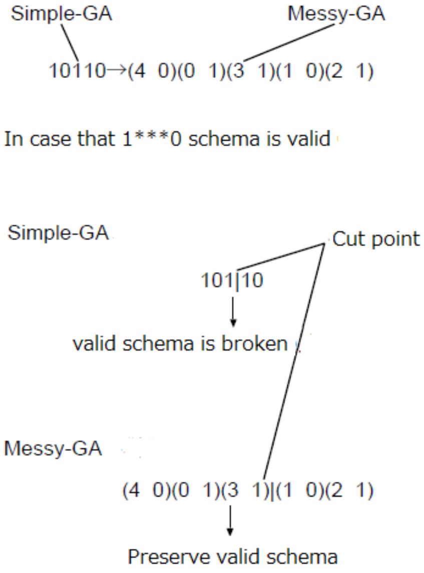


Figure 11 Reason for which the proposed Messy GA clustering is superior to the Simple GA clustering

TABLE II. CLUSTERING PERFORMANCE

Method	Converged iteration number	Number of correct clustering pixels	Percent Correct Clustering: PCC
Simple GA	29238	335	32.71
Messy GA	2998	543	53.03

From Table 2, the converged iteration number of the Simple GA clustering is greater than that of the Messy GA clustering. It takes ten time of processing time is required for the Simple GA clustering in comparison to that of the Messy GA clustering. Meanwhile, the number of correct clustering pixels of the Simple GA clustering is less than that of the Messy GA clustering. This is same thing for percent correct clustering. Therefore, it may say that the proposed Messy GA clustering is superior to the conventional Simple GA clustering. The reason for improvement of convergence performance is that useful building block of chromosome can be created in the initial phase. Through iteration, such useful schema can be preserved and updated efficiently and effectively. On the other hands, relatively long useful schema is broken during the cross over processes for the Simple GA clustering. Therefore, Percent Correct Clustering: PCC of the Simple GA clustering is not good enough. In contrast, the probability of schema broken is reduced for the Messy GA clustering. Figure 11 shows an example for such mechanism. If the cut point for cross over process is situated at the last two digits of the chromosome, and if the 1***0 of chromosome is useful and valid, then valid chromosome of 10110 is broken by cross over for the Simple GA clustering while it is maintained or preserved for the Messy GA clustering.

IV. CONCLUSION

Clustering method for remote sensing satellite image classification based on Messy Genetic Algorithm: GA is proposed. Through simulation study and experiments with real remote sensing satellite images, the proposed method is validated in comparison to the conventional simple GA. It is also found that the proposed clustering method is useful for preprocessing of the classifications.

The converged iteration number of the Simple GA clustering is greater than that of the Messy GA clustering. It takes ten time of processing time is required for the Simple GA clustering in comparison to that of the Messy GA clustering. Meanwhile, the number of correct clustering pixels of the Simple GA clustering is less than that of the Messy GA clustering.

ACKNOWLEDGMENT

The author would like to thank Mr. Koichi Tateno for his effort to conduct simulation and experiments.

REFERENCES

[1] M. Takagi and H.Shimoda Edt. K.Arai et al., Image Analysis Handbook, The University of Tokyo Publishing Co. Ltd., 2001.

[2] A.Yoshizawa, K.Arai, GA based clustering utilizing spectral and spatial context information, Journal of Image and Electronics Engineering Society of Japan, Vol.31, No.2, 202-209, 2003.

[3] K. Arai, A. Yoshizawa, K. Tateno, Messy GA algorith utilizing clustering method for remote sensing satellite inmage clustering, Journal

of Japan Society for Photogrammetry and Remote Sensing, Vol.41, No.5, 34－41, 2003.

[4] Kohei Arai, Learning processes of image clustering method with density maps derived from Self-Organizing Mapping(SOM), ournal of Japan Society for Photogrammetry and Remote Sensing, Vol.43、5、62-67 (2004)

[5] Kohei Arai, Non-linear merge and split method for image clustering, ournal of Japan Society for Photogrammetry and Remote Sensing, Vol.43、5、68-73 (2004)

[6] K. Arai, Quang Bu, GA based ISODATA clustering taking into account concave shape of probability density function, ournal of Japan Society for Photogrammetry and Remote Sensing, Vol.47、1、17一25、2008

[7] K.Arai, Quang Bu, Pursuit reinforcement competitive learning based online clustering and its application of image portion retrievals, Journal of Image and Electronics Engineering Society of Japan, Vol.39,3,301-309,2010

[8] K.Arai, Quang Bu, Accerelation of convergence of the pursuit reinforcement competitive learning based online clustering with learning automaton and its application of evacuatio simulations, Journal of Image and Electronics Engineering Society of Japan, Vol.40, 2, 361-168, 2011.

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 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 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 was appointed professor at Department of Information Science, Saga University in 1990. He was appointed 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 also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission “A” of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers.