Comparative Study between the Proposed GA Based ISODAT Clustering and the Conventional Clustering Methods

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

***Abstract*— A method of GA: Genetic Algorithm based ISODATA clustering is proposed.GA clustering is now widely available. One of the problems for GA clustering is a poor clustering performance due to the assumption that clusters are represented as convex functions. Well known ISODATA clustering has parameters of threshold for merge and split. The parameters have to be determined without any assumption (convex functions). In order to determine the parameters, GA is utilized. Through comparatives studies between with and without parameter estimation with GA utilizing well known UCI Repository data clustering performance evaluation, it is found that the proposed method is superior to the original ISODATA and also the other conventional clustering methods.**

Keywords- GA; ISODATA; Optimization; Clustering.

I. INTRODUCTION

Clustering is the method of collecting the comrades of each-other likeness, making a group based on the similarity and dissimilarity nature between object individuals, and classifying an object in the heterogeneous object of a thing [1]. The classified group calls it a cluster. The criteria which measure how many objects are alike have the degree (similarity) of similar, and the degree (dissimilarity) of dissimilarity [2]. The object with high similarity is one where a value is larger more alike like a correlation coefficient in the degree of similar, and the object with low similarity is not one where the value of the degree of dissimilarity is conversely larger] alike. The degree of dissimilarity is well used in these both. The degree of dissimilarity is also called distance (distance). There is a definition of the distance currently used by clustering how many. The clustering method can be divided into the hierarchical clustering method and the un-hierarchical clustering method [3].

Hierarchical clustering [4] (hierarchical clustering method) is the clustering method for searching for the configurationally structure which can be expressed with a tree diagram or a dendrogram [5], and is method into which it has developed from the taxonomy in biology. A hierarchy method has a shortest distance method, the longest distance method, the median method, a center-of gravity method, a group means method, the Ward method, etc [6]. By a hierarchy method, there are faults, such as the chain effect that computational complexity is large.

A non-hierarchy method is the method of rearranging the member of a cluster little by little and asking for the better cluster from the initial state [7],[8],[9]. It is more uniform than this as much as possible within a cluster, and it is a target to make it a classification which differs as much as possible between clusters. The typical method of a non-hierarchy method has the K-means method and the ISODATA method [10].

A method of GA: Genetic Algorithm [11] based ISODATA clustering is proposed. GA clustering is now widely available. One of the problems for GA clustering is a poor clustering performance due to the assumption that clusters are represented as convex functions. Well known ISODATA clustering has parameters of threshold for merge and split [12],[13]. The parameters have to be determined without any assumption (convex functions). In order to determine the parameters, GA is utilized. Through comparatives studies between with and without parameter estimation with GA utilizing well known UCI Repository data clustering performance evaluation, it is found that the proposed method is superior to the original ISODATA. ISODATA based clustering with GA is proposed in the previous paper [14]. In this paper, comparative study of the proposed ISODATA GA clustering method with the conventional clustering methods is described.

In the next section, theoretical backgrounds on the widely used conventional clustering methods and Genetic Algorithm: GA[1] is reviewed followed by the proposed clustering method based on ISODAT with GA. Then experimental result with simulation data of concave shaped distribution of data is shown for demonstration of effectiveness of the proposed method followed by experimental results with UCI repository[2] of standard datasets for machine learning. In particular, clustering performance of the proposed GA based ISODATA clustering method is compared to those of the other conventional clustering methods. Finally, conclusion and some discussions are described.

Theoretical Background

[1] http://www2.tku.edu.tw/~tkjse/8-2/8-2-4.pdf
[2] http://archive.ics.uci.edu/ml/support/Iris

A. K-Means Clustering

The k-mean method is that of the non-hierarchical type clustering method proposed by MacQueen, Anderberg [15], Forgy and others [16] in the 1960s. Based on the given initial cluster center of gravity, this method uses the average of a cluster and classifies. The process flow is shown as follows.

1. Several k of a cluster is determined and the k cluster center of gravity is given as initial value. There are the following methods in selection of the initial cluster center of gravity. (1) Use the result of the clustering performed before. (2) Presume from knowledge other than clustering. (3) Generate at random.
2. To all individuals, distance with the k cluster center of gravity is calculated, and distance arranges an individual to the cluster used as the minimum.
3. The center of gravity of each cluster is re-calculated by the individual rearranged by 2.
4. If it is below threshold with the number of the individuals which changed the affiliation cluster, it will be regarded as convergence and processing will be ended. When other, it returns to 2. and processing is repeated.

Like this fault, the sum in a cluster which is all the distance of an individual and its cluster center of gravity decreases in monotone. That is, the k-means method is a kind of the climbing-a-mountain method. Therefore, although the k-means method guarantees local optimal nature, global optimal nature is not guaranteed. The result of clustering changes with setup of the initial cluster center of gravity.

B. ISODATA

The ISODATA method is the method developed by Ball, Hall and others in the 1960s. The ISODATA method is a method which added division of a cluster, and processing of fusion to the k-means method. The individual density of a cluster is controllable by performing division and fusion to the cluster generated from the k-means method. The individual in a cluster divides past [a detached building] and its cluster, and the distance between clusters unites them with past close. The parameter which set up division and fusion beforehand determines. The procedure of the ISODATA method is shown as follows.

1. Parameters, such as the number of the last clusters, a convergence condition of rearrangement, judgment conditions of a minute cluster, branch condition of division and fusion, and end conditions, are determined.
2. The initial cluster center of gravity is selected.
3. Based on the convergence condition of rearrangement, an individual is rearranged in the way of the k-means method.
4. It considers with a minute cluster that it is below threshold with the number of individuals of a cluster, and accepts from future clustering.
5. When it is more than the threshold that exists within fixed limits which the number of clusters centers on the number of the last clusters, and has the minimum of the distance between the cluster centers of gravity and is below threshold with the maximum of distribution in a cluster, clustering regards it as convergence and ends processing. When not converging, it progresses to the following step.
6. If the number of clusters exceeds the fixed range, when large, a cluster is divided, and when small, it will unite. It divides, if the number of times of a repetition is odd when there is the number of clusters within fixed limits, and if the number is even, it unites. If division and fusion finish, it will return to 3. and processing will be repeated.

Division of a cluster: If it is more than threshold with distribution of a cluster, carry out the cluster along with the first principal component for 2 minutes, and search for the new cluster center of gravity. Distribution of a cluster is re-calculated, and division is continued until it becomes below threshold.

Fusion of a cluster: If it is below threshold with the minimum of the distance between the cluster centers of gravity, unite the cluster pair and search for the new cluster center of gravity. The distance between the cluster center of gravity is re-calculated, and fusion is continued until the minimum becomes more than threshold.

Although the ISODATA method can adjust the number of certain within the limits clusters, and the homogeneity of a cluster by division and fusion, global optimal nature cannot be guaranteed. Since the ISODATA method has more parameters than the k-means method, adjustment of the parameter is still more difficult.

C. Heredity Algorithm

A heredity algorithm (Genetic Algorithms: GA) is an optimization algorithm modeled after the theory of evolution of Darwin, and it will be advocated by Holland[3] in the 1960s. The solution in question is expressed as an individual and an each object is constituted from GA by the chromosome. An individual evolves by selection, intersection, and mutation, and searches for an optimum solution.

The general procedure of GA is shown as follows.

1. N individuals with a chromosome are generated as the initial population (population). Simultaneous search of the N points can be carried out by these N individuals.
2. Fitness value is searched for based on the fitness value function beforehand defined to each individual.
3. Selection is performed based on fitness value. That is what is screened out of N individuals of current generation and the thing which survives the next generation. The probability of surviving the next generation becomes high so that the fitness value of an individual is high, but the low individual of fitness value may also survive in the next generation. This is a role which controls lapsing into a partial solution.

[3] http://www2.econ.iastate.edu/tesfatsi/holland.gaintro.htm

Tournament selection[4]: It is the method of repeating this process until it selects a certain number of individuals at random from the population, fitness value chooses the best thing in it and the population's number of individuals is obtained.

Elite strategy[5]: How many individuals with the maximum fitness value call it the elite. The method of certainly leaving the elite to the next generation regardless of a selection rule is called elite strategy. The elite individual saved by an elite strategy participates in neither intersection nor mutation.

4. By the set-up intersection probability or the intersection method, the selected individual is crossed (crossover) and a new individual is generated.
5. By the method of the set-up mutation rate or mutation, mutation is performed and a new individual is generated.
6. Fitness value is re-calculated to a new chromosome group.
7. If end conditions are fulfilled, let the best individual then obtained be the semi optimum solution in question. Otherwise, it returns to 3.

GA is the multipoint search method, and is excellent in global searching ability, also is widely applied to various optimizations or a search problem.

D. Real Numerical Value GA

Early GA performed intersection and mutation by the bit string which carried out the binary coding of the variable, and has disregarded the continuity of a variable. On the other hand, GA which performs intersection in consideration of the continuity of a variable and mutation is called the real numerical value GA (Real-Coded Genetic Algorithms)[6] using the numerical value itself. In this research, the threshold of an initial cluster center and division/fusion is optimized based on the real numerical value GA.

GA with a general flow of processing of the real numerical value GA is the same. Since the coding method is merely different, the original intersection method and the mutation method are used.

The intersection method of real numerical value GA daily use has the BLX-alpha method[7], single modal normal distribution crossing method (Uni-modal Normal Distribution crossover: UNDX)[8]:, etc., and the mutation method has mutation, uniform mutation, etc. by a normal distribution.

The BLX-alpha crossing method: This intersection method determines a child as follows,

1. Two parent individuals are set to a and b.
2. The section [A, B] of intersection is calculated by the following formulas.

$$A = \min(a,b) - \alpha|a-b| \qquad B = \max(a,b) + \alpha|a-b| \qquad (1)$$

3. A uniform random number determines a child individual from the section [A, B].

Mutation by a normal distribution: It can be happened mutation by a normal distribution. The normal distribution used at this time will be decided with the random number according to the normal distribution of the average of x distribution delta 2, if a parent individual is set to x. The individual generated exceeding the range of x [XMIN, XMAX] is stored in within the limits.

II. Proposed Clustering Method

It decided to use GA also for the determination of the threshold of the separation in clustering by ISODATA, and integration. It is because a clustering result will constitute inevitably the cluster that cluster distribution becomes the best for a case to a convex function wholly in the bottom if this sets up an fitness value function which makes the maximum the ratio of synthesis of distribution between clusters, and synthesis of cluster internal variance. By the method of repeating separation and integration like ISODATA, it decided to avoid an above-mentioned problem by controlling this threshold. The consecutiveness distribution form based on it needs the concept of the distance in the feature space, and to be judged for this control, and in order to perform this, GA is used in this paper.

A. Partial Mean Distance

As partial mean distance is shown in Fig.1, the average of the distance between the individuals belonging to the same cluster of a certain part within the limits is called partial mean distance. It can ask for the sum of partial mean distance all over the districts by moving the range of a part by the Moving Window[9] method little by little. The window of the Moving Window method here is a super-sphere in n-dimensional Euclidean space.

Since distribution of a cluster is not necessarily uniform distribution, when the window of the Moving Window method

[4] http://www.google.co.jp/#hl=ja&rlz=1W1GGLD_ja&q=tonament+selection+genetic+algorithm&oq=tonament+selection+genetic+algorithm&aq=f&aqi=&aql=&gs_l=serp.3...4859.11766.0.13031.25.24.0.0.0.6.344.4800.0j14j8j2.24.0...0.0.EzWVG3xcR3I&pbx=1&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=48edeecabd799c01&biw=1280&bih=799

[5] http://www.sersc.org/journals/IJSH/vol2_no2_2008/IJSH-Vol.2-No.2-W03-Solving%20Unbounded%20Knapsack%20Problem.pdf

[6] http://www.google.co.jp/#hl=ja&site=&source=hp&q=real+coded+genetic+algorithm&rlz=1W1GGLD_ja&oq=real+coded+&aq=0L&aqi=g-L6&aql=&gs_l=hp.1.0.0i19l6.2438.7016.0.11266.11.9.0.2.2.0.250.1454.0j8j1.9.0...0.0.AZfbfrGsWhw&pbx=1&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=ede0ef6aa5da214e&biw=1280&bih=758

[7] http://www.iitk.ac.in/kangal/resources.shtml

[8] http://www.google.co.jp/#hl=ja&rlz=1W1GGLD_ja&q=unimodal+normal+distribution+crossover+genetic+algorithm&oq=unimodal+normal+distribution+crossover+genetic+algorithm&aq=f&aqi=&aql=&gs_l=serp.3...4828.22188.0.23157.52.44.0.0.0.10.250.6675.0j38j6.44.0...0.0.n-0BTKOtkRQ&pbx=1&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=48edeecabd799c01&biw=1280&bih=758

[9] http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=548342&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel3%2F3974%2F11463%2F00548342.pdf%3Farnumber%3D548342

is moved at equal intervals, useless calculation may be carried out in a place without an individual. In order to avoid this, in this paper, in all individuals, it will move for every individual and the sum of partial mean distance all over the districts will ask for the super-sphere centering on an individual.

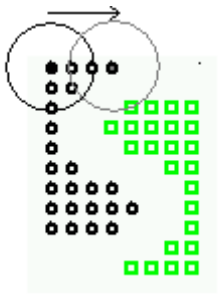

Figure 1Partial mean distance

Making the sum of partial mean distance all over the districts into the minimum, the density of an individual, an individual can be made to belong to a separate cluster along a crevice in a small place, i.e., a place with a crevice that is, the boundary line of a cluster can be made so much to a concave set.

B. Difference from the Conventional ISODATA Method

The ISODATA method is a method which cluster distribution assumes to be a convex function. When cluster distribution is a concave function, by the ISODATA method, it can respond to some extent by division and fusion, but if the procedure of the conventional ISODATA method is followed, the cluster classified correctly once may be destroyed.

Since equivalence will be carried out if the individual rearrangement in the process of the ISODATA method is the k-means method in fact and a cluster can be divided in a straight line with a Voronoi figure[10], when cluster distribution is a concave function, the cluster divided by division and fusion with the curve may be destroyed by rearrangement of an individual. Then, after the proposal method unites the last of the ISODATA method, it does not rearrange an individual and is ended.

When cluster distribution is a concave function, suppose that former data was divided by the threshold of suitable division. Since the distance between clusters changes into the united process when uniting a cluster after this, as for the turn of fusion, a result will be affected. When it does so, even if there is a threshold of suitable fusion, a desirable fusion result may not be brought. By the proposal method, in order to depend for the fusion result of a cluster only on the threshold of fusion, simultaneous fusion of the cluster filled to the threshold of fusion is carried out.

Moreover, since the center of a cluster is presumed by Real Coded GA: RCGA by the proposal method, even if a clustering result reduces the number of times of a repetition of the ISODATA method for which it does not depend on correction of the center of a cluster by repetition of the ISODATA method to some extent, it hardly influences a clustering result. Therefore, in this paper, the number of times of a repetition of the ISODATA method is set to 2 for the improvement in calculation speed.

C. Selection of Fitness Value Function

Since the cluster that *F* is made into the maximum and that cluster distribution will become [as for the result of clustering] the best for a case to a convex function wholly in the bottom if an fitness value function setup is carried out will be constituted, it is not suitable when cluster distribution is a concave function.

As make into the minimum of the sum of partial mean distance all over the districts, since only the crevice within the limits between parts will be observed if a fitness value function setup is carried out, a cluster may not become a lump.

$$\text{Fitness} = F + \frac{m}{d} \qquad (2)$$

In this equation, *F* expresses a false F value, d denotes the sum of partial mean distance all over the districts, and m expresses weight.

Here, when asking for the sum of partial mean distance, selection of the range of a part has large influence on a result. If the range of a part is too small, a crevice cannot be covered and the boundary line of a cluster cannot be made correctly. Moreover, when the range of a part is too large, locality may lose. The radius of a super-sphere which expresses the range of a part with the proposal method for an object as one cluster is enlarged little by little from the shortest distance between individuals to the maximum distance, and it asks for the sum of partial mean distance. In the time of the radius of a super-sphere becoming at least in the width of a crevice, the sum of partial mean distance reaches one peak. In order to carry out the certain cover of the crevice, the sum of partial mean distance makes a few radiuses this becomes a peak, the range of the part actually using a super-sphere with a large radius.

D. Set-up Parametersfor RCGA

The selection method, tournament selection and an elite strategy is used. The size of tournament selection is set as 3.

Using the BLX-alpha method, the intersection method sets the value of alpha as 0.5, and sets up intersection probability to 70%.

Using the mutation method by a normal distribution, the mutation method sets the value of sigma as 0.5, and sets up mutation probability to 1%.

Termination conditions: the elite, five-generation maintaining t as a thing and five generations of differences of the average fitness values and the elite's fitness value continuing 2% in within the limits.

By the ISODATA method, the threshold of an initial cluster center, division, and fusion is presumed by GA.

III. EXPERIMENTS

A. Performance Evaluation Method

A different clustering result is obtained from the separate clustering method in many cases. Also by the same clustering

[10] http://otago.ourarchive.ac.nz/handle/10523/765

method, it sometimes often results in changing with setup of a parameter. The criteria of evaluation are needed in order to compare the result of clustering. *F* value (pseudo F statistic) is one of the valuation bases often used. *F* is defined by the following formulas.

$$F = \frac{\frac{\sum_{i=1}^{n}(l_i - \bar{l})^2 - \sum_{j=1}^{k}\sum_{i \in C_j}(l_i - \bar{l}_j)^2}{k-1}}{\frac{\sum_{j=1}^{k}\sum_{i \in C_j}(l_i - \bar{l}_j)^2}{n-k}} \quad (3)$$

where *n* as for the total number of individuals and *k*, as for *Cj*, Clusters j and *li* express the number of clusters, Individual i and #j express the average of Cluster j, and # expresses the average of all individuals. *F* is criteria which consider simultaneously the variation within a cluster, the variation between clusters, and the number of clusters, and the figure is the ratio of distribution between groups, and group internal variance. Since group internal variance with large distribution between groups means the small thing if *F* is large, a clustering result shows a good thing. However, since cluster distribution assumes it as the convex function in *F*, in the case of the concave function, it is not suitable.

In the case where the correct answer of a classification is known, the error *E* of a result can be searched for from the number of individuals c classified correctly.

$$E = \frac{n-c}{n} \times 100\% \quad (4)$$

B. Selection of Fitness Value Function

The proposal method experiments by making the data of simple degree of convection to verify whether it can respond not only when cluster distribution is a convex function, but in the case of a concave function.

As shown in Fig.2, it experiments using the data containing two clusters of degree of convection. It clusters by the ISODATA method and the proposal method with a random parameter, and the result of clustering is compared. The result of an experiment is shown like a lower figure. The error which cannot understand the cluster of degree of convection in a straight line, and cannot classify it according to the conventional ISODATA method correctly from this experimental result is 12.5%. And by the proposal method, it turns out that an error becomes 0% and it can classify according to division and fusion correctly with a curve.

C. Experiemnt 2

Next, it experiments using the Iris data set of UCI repository[11], a Wine data set, a Ruspini data set, and a New thyroid data set. Iris is a 4-dimensional data set with a number of individuals 150 and three categories. Wine is a 13-dimensional data set with a number of individuals 178 and three categories. Ruspini is a 2-dimensional data set with a number of individuals 75 and four categories. New thyroid is a 5-dimensional data set with a number of individuals 215 and three categories. These four data sets are criteria data sets often used for comparison of the clustering method. When clustering an Iris data set by the case where a parameter is presumed by GA, change of the fitness value of 50 generations, i.e., the process of convergence, is shown in Fig. 5.

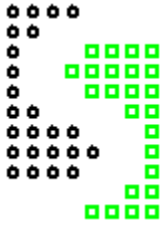

Figure 2 Original data

Figure .3 ISODATA method (Random)

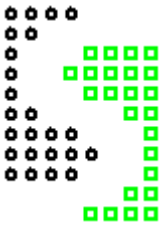

Figure 4 Proposed method (ISODATA-GA)

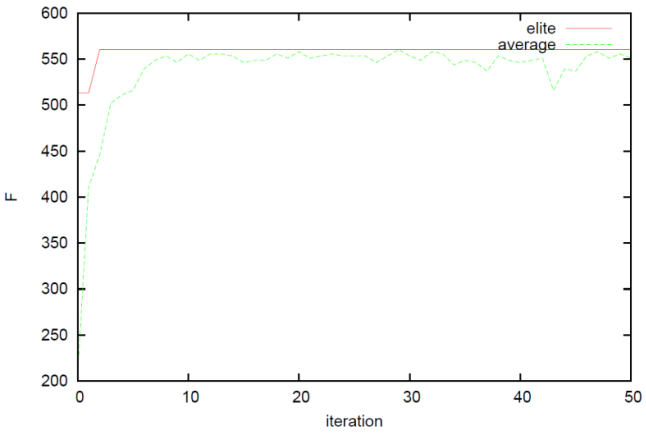

Figure 5 Convergence process of GA

In this figure, a red line expresses the elite's fitness value, and expresses the average of fitness value with the green line. This figure shows being completed by the average of fitness value. Then, the similar result was obtained in the experiment of a Wine data set, a Ruspini data set, and a new thyroid data set. The data of the experimental result of these data sets is gathered in Table 1.

TABLE I. CLUSTERING PERFORMANCE WHEN THE NUMBER OF CLUSTER IN AGREEMENT WITH TEH NUMBER OF TARGET CLUSTERS

Error (%)	The ISODATA method (Random)	The proposed method
Iris	21.33	11.33
Wine	34.65	11.34
Ruspini	46.76	4.00
New thyroid	31.63	15.81

[11] http://archive.ics.uci.edu/ml/(standard dataset for machine learning)

In the table, compared with the conventional ISODATA method, the direction of the proposal method shows that the error decreases 22.97%, respectively. Although optimizing took about 100-time long time from the conventional ISODATA method, accuracy went up certainly.

D. Experiemnt 3

Comparative study on clustering performance and processing time between the proposed ISODAT GA method and the conventional clustering methods of (1) Minimum Distance Method, (2) Maximum Distance Method, (3) Median Method, (4) Gravity Method, (5) Group Average Method, (6) Ward Method, (7) K-Means Method with the best random initial cluster center selection, (8) K-Means Method with the average random initial cluster center selection, (9) ISODATA Method with the best random initial cluster center selection, (10) ISODATA Method with the average random initial cluster center selection is conducted with Iris dataset of UCI repository. Table 2 shows clustering performance of the proposed and the other conventional clustering methods.

Although processing time required for the proposed ISODATA GA clustering is three times much longer, F value which represents the ratio of inner cluster variance to between cluster variance of the proposed ISODATA GA shows the best value together with clustering error (it also shows almost minimum value). In particular, 6.2% of clustering error is reduced by the proposed ISODATA GA clustering method in comparison to the conventional ISODAT with best selection of initial cluster center randomly (ISODATA(Ran.Best)), separability between clusters, *F* are almost same though. Also 68.7% of separability improvement is observed in comparison to the conventional ISODATA(Ran.Ave.). Therefore, parameter (merge and split of the clusters) estimation based on GA as well as initial cluster center determination with GA are effective to improve clustering performance.

Iris dataset is four dimensional data which consists of 150 of the number of data points with the number of categories (cluster) of three. It may say that Iris is the typical dataset among UCI repository.

V. Conclusion

The proposed clustering method is based on the conventional ISODAT method. One of the problems of the ISODATA method is relatively poor clustering performance. For instance, ISODAT as well as the other conventional clustering method do not work well if the probability density function of data is distributed as concave, then linear discrimination function does not work well. Taking such probability density function into account, parameters for merge and split of the clusters can be adjusted with GA. Also the proposed ISODATA GA method determines initial cluster center with GA. Clustering performance depends on the designated initial cluster center. Therefore, if not appropriate initial cluster center is determined, then cluster results become bad. The proposed ISODATA GA method determines most appropriate initial cluster center by using GA. Therefore, cluster results become excellent. The experimental results show that most appropriate cluster result can be obtained with the proposed ISODATA GA for the situation of shape independent clustering with concave shape of input data distribution. Also the experimental results with UCI repository show that the proposed ISODATA GA method is superior to the conventional ISODATA clustering method with randomly determined initial cluster center. It is also found that the proposed ISODATA GA method is superior to the other typical conventional clustering methods.

TABLE 2 Clustering Performance of the Proposed ISODATA with GA for IRIS Dataset of UCI Repository.

	F	*E*(%)	Process Time(s)
Minimum Distance Method[12]	277.493	32	0.14
Maximum Distance Method[13]	484.899	16	0.14
Median Method[14]	501.303	10	0.156
Gravity Method[15]	555.666	9.33	0.156
Group Average Method[16]	399.951	25.33	0.14
Ward Method[17]	556.841	10.67	0.14
K-means(Ran.Best)[18]	560.366	11.33	0.06
K-means(Ran.Ave.)[19]	210.279	46.23	0.06
K-means(GA)[20]	560.4	10.67	0.225
ISODATA(Ran,Best)[21]	560.366	11.33	0.313
ISODATA(Ran,Ave.)[22]	175.465	38.64	0.313
ISODATA(GA)[23]	560.4	10.67	1.523

Acknowledgment

The author would like to thank Dr. XingQiang Bu for his effort to experimental studies.

References

[1] Kohei Arai, Fundamental theory for pattern recognition, Gakujutu-Tosho-Shuppan Pub. Co., Ltd.,1999.

[2] Hartigan, J.A., Clustering Algorithms, NY: Wiley, 1975.

[3] Anderberg, M.R. , Cluster Analysis for Applications, New York: Academic Press, Inc., 1973.

[4] Bottou, L., and Bengio, Y., "Convergence properties of the K-means algorithms," in Tesauro, G., Touretzky, D., and Leen, T., (eds.) Advances inNeural Information Processing Systems 7, Cambridge, MA: The MIT Press, 1995

[5] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.

[12] http://en.wikipedia.org/wiki/Cluster_analysis
[13] http://en.wikipedia.org/wiki/Hierarchical_clustering
[14] http://www.cra.org/Activities/craw_archive/dmp/awards/2003/Mower/KMED.html
[15] http://dl.acm.org/citation.cfm?id=585147.585174
[16] http://nlp.stanford.edu/IR-book/html/htmledition/group-average-agglomerative-clustering-1.html
[17] http://www.statsoft.com/textbook/cluster-analysis/
[18] http://en.wikipedia.org/wiki/K-means_clustering
[19] http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
[20] http://airccse.org/journal/ijaia/papers/0410ijaia3.pdf
[21] http://www.cs.umd.edu/~mount/Projects/ISODATA/
[22] http://www.yale.edu/ceo/Projects/swap/landcover/Unsupervised_classification.htm
[23] http://www.yale.edu/ceo/Projects/swap/landcover/Unsupervised_classification.htm

[6] L.Breiman and J.Friedman, Predicting multivariate responses in multiple linear regression, Technical report, Department of Statistics, University of California, Berkley, 1994.

[7] R.J.Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine Learning, 8, 3-4, 229-256, 1992.

[8] L.P. Rieber, Computer, Graphics and Learning, Madison, Wisconsin: Brown & Benchmark, 1994.

[9] C. Yi-tsuu, Interactive Pattern Recognition, Marcel Dekker Inc., New York and Basel, 1978.

[10] R.K. Pearson, T. Zylkin, J.S. Schwaber, G.E. Gonye, Quantitative evaluation of clustering results using computational negative controls, Proc. 2004 SIAM International Conference on Data Mining, Lake Buena Vista, Florida, 2004.

[11] Goldberg D., Genetic Algorithms, Addison Wesley, 1988, or, Holland J.H., Adaptation in natural and artificial system, Ann Arbor, The University of Michigan Press, 1975.

[12] Trevor Hastie, Robert Tibshirani, Jerome Friedman The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.

[13] Jensen, J.R., Introductory Digital Image Processing. Prentice Hall, New York, 1996.

[14] Kohei Arai and XianQiang Bu, ISODATA clustering with parameter (threshold for merge and split) estimation based on GA: Genetic Algorithm, Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 17-23, 2007

[15] MacQueen, J.B., Some Methods for Classification and Analysis of Multivariate Observations,, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297,1967.

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.