'93 KACC (1993.10.20~22)

Application of Genetic Algorithms to Cluster Analysis

Takanori Tagami, Sadaaki Miyamoto and Yoshio Mogami
Department of Information Science and Intelligent Systems
Faculty of Engineering, University of Tokushima
Minamijosanjima 2-1, Tokushima 770, JAPAN

Abstract

The aim of the present paper is to show the effectiveness of Genetic Algorithm for data classification problems in which the classification criteria are not the Euclidean distance. In particular, in order to improve a search performance of Genetic Algorithm, we introduce a concept of the degree of population diversity, and propose construction of genetic operators and the method of calculation for the fitness of an individual using the degree of population diversity. Then, we investigate their performances through numerical simulations.

1. Introduction

Genetic Algorithm (GA) is known as a most effective method to solve combinatorial optimization problems by simulating the process of natural evolution and natural genetics, and many theoretical studies and its applications have been considered [1)~4)]. In this paper, the procedure of GA is applied to cluster analysis.

Cluster analysis or clustering is a basic technique in the field of data analysis. Cluster analysis may be used to reveal categorical structures in the data. It is applied to a variety of scientific data classification problems such as those in life sciences, medicine, engineering and so on. Then, many clustering methods are already proposed [5)~8)]. Most of these clustering methods deal with the clustering in which the classification criteria are the Euclidean distance, for example, its typical method is MacQueen's k-means method [8)]. Applications of GA to the clustering based on the Euclidean distance have already been reported [9)~11)]. However, the clustering not based on the Euclidean distance have a larger number of applications when compared with the clustering based on the Euclidean distance. So far, in the case of the non-Euclidean distance case, the MST (Minimum Spanning Tree) method have been applied to the clustering, but the MST method has a problem, i.e., sensitivity for an observation noise of data. Thus, a practically useful method for the non-Euclidean distance has not been proposed yet. This paper intends an investigation of the effectiveness of a combinatorial optimization method using GA for clustering, and is concerned especially with the clustering not based on the Euclidean distance.

In a traditional GA, its dispersion of the search performance will be large, depending on the operators using random numbers or the control parameters. In this paper, for the purpose of reducing the dispersion of the search performance, a concept of the degree of population diversity is introduced as an index for an internal state of the whole population, and we use this index for control parameters of the genetic operators such as crossover, mutation and selection and the method of calculation for the fitness of an individual.

2. Construction of GA for clustering

2.1 Outline of GA

For the purpose of solving a given problems, a GA requires a feasible solution for the problem to be coded as a finite length strings using some numbers or alphabets. This operation is called *genetic coding*, and this string is called *individual*. The position of the string is called *locus*. The variable at a locus is called *gene*, and its value *allele*. A collection of a member of individuals is used for solving problems. This collection is called a *population*. The *fitness* of each individual is evaluated for a given problem, and individuals of the popu-

lation are gradually improved by using the fitness and genetic operators such as crossover, mutation and selection.

A general procedure of GA consists of the following steps:

1. Initialize the genes of each individual in the population $P(t=0)$.
2. Generate $P(t+1)$ from $P(t)$ as follows: evaluate fitness of each individual in $P(t)$; select individuals from $P(t)$ using fitness; recombine them using genetic operators;
3. If time is up, stop and return the best individual; if not, set $t = t+1$ and go to 2.

An iteration is called a *generation*. The index t indicates the number of generations.

2.2 Clustering

Clustering is a tool for exploring the structure of the data that does not require the assumptions common to most statistical methods. It is called *unsupervised learning* in the literture of pattern recognition and artificial intelligence [6].

To begin with, we will introduce several symbols. A set of objects which should be divided, is denoted by $A = \{a_1, a_2, \cdots, a_n\}$, and the elements $a_1, a_2, \cdots, a_n$ are called "objects", or simply called "data". The set A should be divided into clusters C_i $(i = 1, \cdots, K)$. Clusters mean a family of disjoint subsets whose union coincides with the data set. $(C_1 \cup C_2 \cup \cdots \cup C_K = A$ and $C_i \cap C_j = \emptyset)$

Clustering requires a classification criterion that means an index of alikeness or association between pairs of data. In this paper, it is called *dissimilarity*. The dissimilarity between a_i and a_j data is denoted $d(a_i, a_j)$ and must satisfy the following three properties:

$$\begin{aligned} &1)\ d(a_i,a_j) \geqq 0 \\ &2)\ d(a_i,a_j) = 0 \Leftarrow i = j \\ &3)\ d(a_i,a_j) = d(a_j,a_i) \end{aligned} \qquad (1)$$

The degree of relation between a_i and a_j becomes small when the dissimilarity $d(a_i, a_j)$ is large. Thus, the aim of the clustering is to sort the data into clusters such that the dissimilarity is low among members of the same cluster and high between members of different clusters.

In particular, the clustering problem leads to the minimization of a function D[12]:

$$\begin{aligned} D &= \frac{1}{2}\sum_{k=1}^{K} n_k s_k \\ s_k &= \frac{1}{n_k^2}\sum_{a_i \in C_k}\sum_{a_j \in C_k} d(a_i, a_j) \end{aligned} \qquad (2)$$

where n_k denote the number of data which are included in the cluster C_k. The parameter K indicates the number of clusters.

When the GA is applied to the clustering problem, we have to consider three issues: (1) genetic coding for the individual; (2) construction of genetic operators (crossover, mutation, selection); (3) calculation for the fitness of individual.

2.3 Genetic coding for the individual

Suppose that a clustering problem have to divide the data a_i $(i = 1, \cdots, n)$ into K $(K \geqq 2)$ clusters. Then, the genes x_i $(i = 1, \cdots, n)$ of one individual X are denoted as integers in the interval $[1, K]$. For example, $n = 8$, $K = 3$, individual X which is expressed as below means that $\{a_3, a_4, a_8\}$ belong to the first cluster C_1 , $\{a_2, a_5\}$ belong to the second cluster C_2 and $\{a_1, a_6, a_7\}$ belong to the third cluster C_3, respectively.

data :	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
X :	3	2	1	1	2	3	3	1

That is, these numbers of the individual correspond to the cluster numbers of the data.

3. Degree of population diversity

3.1 Introduction of the degree of diversity

In a GA, operations using the fitness and genetic operators are carried out for the individuals of the population, and the individuals are gradually improved. We observe that with the progress of the generations, a locus is concentrating a specific allele as shown in **Fig.1**. In this case, each allele is equivalent to the cluster number of the data. Then, the population finally consists of individulals having the same arrangement of allele. Accordingly, the population diversity will be decreased.

Observing this, we note the possibility of quantifying the degree of population diversity. If we adequately define the degree of population diversity as an index for an internal state of the whole population, we can use this index for control parameters of the genetic operators for developing a

method of calculation for the fitness of an individual in order to obtain better individuals.

$$P(0):\quad \left[\begin{array}{c|cccccccc} X_1 & 1 & 2 & 3 & 2 & 1 & 2 & 3 & 2 \\ X_2 & 2 & 3 & 1 & 3 & 2 & 3 & 2 & 1 \\ X_3 & 2 & 1 & 2 & 3 & 3 & 1 & 1 & 2 \\ \vdots & & & & & \vdots & & & \\ X_p & 1 & 3 & 1 & 1 & 3 & 3 & 2 & 1 \end{array}\right]$$

$$\Downarrow \text{ after } t \text{ generations}$$

$$P(t):\quad \left[\begin{array}{c|cccccccc} X_1 & 3 & 1 & 2 & 1 & 2 & 3 & 2 & 1 \\ X_2 & 3 & 1 & 1 & 1 & 2 & 1 & 2 & 1 \\ X_3 & 3 & 1 & 2 & 1 & 2 & 2 & 3 & 1 \\ \vdots & & & & & \vdots & & & \\ X_p & 3 & 1 & 3 & 1 & 2 & 3 & 1 & 1 \end{array}\right]$$

Fig.1 An illustration of the decrease of diversity

3.2 Calculation for the degree of diversity

Suppose that a population $P(t)$ consist of p individuals that have n length string. Then, the degree of population diversity $V(t)$ and the degree of diversity for each locus v_k $(k = 1, \cdots, n)$ are defined as follows:

$$P(t):\quad \left[\begin{array}{c|cccc} X_1 & x_{11} & x_{12} & \cdots & x_{1n} \\ X_2 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_p & x_{p1} & x_{p2} & \cdots & x_{pn} \\ \hline & v_1 & v_2 & \cdots & v_n \end{array}\right]$$

$$V(t) = \frac{1}{n}\sum_{k=1}^{n} v_k \qquad (3)$$

$$v_k = \frac{\gamma_k}{\gamma_{\max}} \qquad (k = 1, \cdots, n) \qquad (4)$$

$$\text{where} \quad \left(\begin{array}{rcl} \gamma_{\max} & = & \max\limits_{1 \le k \le n} \gamma_k \\ \gamma_k & = & \sum\limits_{i=1}^{p}\sum\limits_{j=1}^{p} \delta_{ij}^{k} \\ \delta_{ij}^{k} & = & \begin{cases} 1 & \text{if } x_{ik} \neq x_{jk} \\ 0 & \text{otherwise} \end{cases} \end{array}\right)$$

In general, these values are near 1 in the initial population $P(0)$. They are decreasing as the generation proceeds and when the whole population is with the same individuals, these valuses will be 0. Next, we propose the construction of genetic operators and the method of calculation for the fitness which are based on the degree of population diversity.

3.3 Crossover

We consider the formation of mask pattern at uniform crossover using the degree of diversity for each locus.

A traditional uniform crossover has a problem which is inclined to break a good connection at the locus. The reason for this is that the operations at uniform corssover depends on the mask pattern which consists of the bit value with an equal probability. The desirable allele for a given problem concentrate on the loci are with the low degree of diversity, and we generate the mask pattern by connecting loci with the low degrees of diversity regarded as one locus.

Namely, when the degree of diversity at the locus between v_k and v_{k+i} satisfy the following inequality:

$$v_k, v_{k+1}, \cdots, v_{k+i} \quad < \quad \rho_c \qquad (5)$$

the elements of the mask pattern between kth and $(k+i)$th are regarded as one element. The parameter ρ_c is a positive real number indicating a threshold for connecting loci. **Fig.2** indicates an example of uniform crossover using the proposed method.

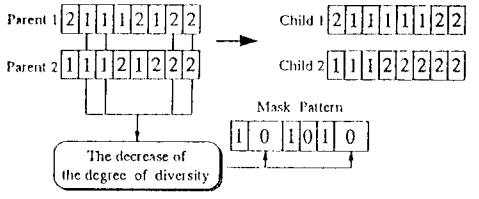

Fig.2 An example of uniform crossover using the proposed method

3.4 Mutation

A traditional mutation is carried out randomly for the whole population according to the mutation rate determined beforehand. The state of loci with the low degree of diversity are frequently involved in the local minimum for a given problem, and we consider the mutation at locus units for escaping from a local minimum.

The mutation rate ρ_m^k of the kth locus is defined as follows:

$$\rho_m^k = \rho_m \times (1.0 - v_k + \alpha) \qquad (0 < \alpha \ll 1) \quad (6)$$

where v_k denotes the degree of diversity on the kth locus and ρ_m indicates the initial mutation rate.

In this mutation, The mutation rate at a locus becomes large when the degree of diversity at the locus is small.

3.5 Selection

We consider the selection which limits the number of individuals having the same arrangement of allele.

First, the maximum number of the same arrangement is defined by an integral number N. Then, choose the individuals which are preserved to the next generation using roulette wheel method and elite method [3]. With the variation of the degree of diversity for the population, the number of the individuals with the same arrangement is defined as follow:

$$\mathrm{int}(N \times V(t)) + 1 \tag{7}$$

The number of the same arrangement becomes small when the degree of populaton diversity $V(t)$ is small.

3.6 Calculation for the fitness

The fitness f_i of individual X_i is defined by using the degree of population diversity $V(t)$ and the evaluation values D_i which is defined for X_i using Eq.(2) as follows:

$$\begin{aligned} f_i &= F_{\max} - F_s \times (O_i - 1) \\ F_s &= \frac{1}{p-1}\left(F_{\max} - F_{\min} - \eta(1 - V(t))\right) \end{aligned} \tag{8}$$

where O_i denotes the order for the evaluation value of the individual X_i in the population ($1 \leqq O_i \leqq p$). Namely, X_i with the highest fitness has $O_i = 1$, X_j with the second highest fitness has $O_j = 2$, and so on. The parameter p denotes the size of population. The parameters $F_{\max}, F_{\min}$ denote the maximum value and the minimum value with the fitness range, respectively. The parameter η indicates a positive real number which determines the minimum fitness $f_{\min}$, and satisfy $0 < \eta \leqq F_{\max} - F_{\min}$.

In this calculation for the fitness, the difference between the maximum fitness value and the minimum fitness value becomes small when the degree of population diversity $V(t)$ is small. That is, as the degree of population diversity is smaller, it is easier to combine an individual of high rank and the one of low rank.

4. Simulation

In the simulations, we investigate the search performances of the traditional GA and the present GA using the proposed operations for a clustering problem. **Table 1** shows the construction of each GA and control parameters. GA1 and GA2 are the traditional search using one-point crossover and uniform crossover, while GA3 uses the search using the proposed operations. **Table 2** indicates the parameters of the experiment.

Table 1 Control parameters for GA1, GA2 and GA3

	crossover	mutation	selection
GA1	one-point crossover	$p_m = 0.01$	roulette wheel, elitism
GA2	uniform crossover		
GA3	proposed method		
	$p_c = 0.01$	$p_m = 0.02$	$N = 2$

Table 2 Parameters for experiment

terminal generation	population size	trial number
100	50	20

4.1 Experiment

Suppose that the clustering problem have to divide 50 data a_i ($i = 1, \cdots, 50$) into 3 clusters C_k ($k = 1, 2, 3$). Each data have two positive real numbers $a_i = (a_{i1}, a_{i2})$ for two attributes $\mathcal{A}_1, \mathcal{A}_2$. **Fig.3** shows the arrangement of each data when the x-axis is for the first attribute value and the y-axis is for the second attribute value.

The dissimilarity between data a_i and a_j is defined by using two attribute values as follows:

$$\begin{aligned} d(a_i, a_j) &= (1.0 - \cos\theta_{ij}) \times 100.0 \\ \cos\theta_{ij} &= \frac{\sum_{k=1}^{2} a_{ik} a_{jk}}{\left(\left[\sum_{k=1}^{2} a_{ik}^2\right]\left[\sum_{k=1}^{2} a_{jk}^2\right]\right)^{1/2}} \end{aligned} \tag{9}$$

Namely, θ_{ij} is the angle between vectors a_i and a_j.

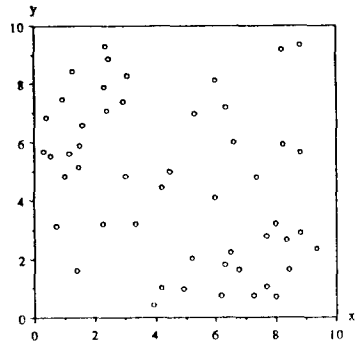

Fig.3 Arrangement of data for clustering

The simulation results are shown in **Fig.4** $\sim$ **7**. **Fig.4** and **Fig.5** show the convergence of the evaluation function and the variations of the degree of population diversity $V(t)$. **Fig.6** shows the convergence of the evaluation function without mutation. **Fig.7** shows the variances of the evaluation value at each generation. These results show the average of 20 trials for each GA.

From these results, we can summarize as follows:

(a) The GA3 using the concept of the degree of population diversity is superior in the convergence of the evaluation function to other GA, and maintains the population diversity comparatively long.

(b) The performances of the GA1 and GA2 are degraded by the effects of mutation when compared with the GA3.

(c) The GA3 reduces the variances of evaluation function at each generation. After 100 generations, the variance is zero. Namely, the GA3 leads to the identical individual at all trials.

Next, we compare the performances of the GA3 and the MST method. We consider the clustering when the attribute values of the data at the **Fig.3** have observation noises. The classification results are shown in **Fig.8, 9**. In these figures, the same marks are classified as members of the same cluster.

These results can be concluded the following:

(a) The MST method is indicated sensitive responses for the noise, while the GA3 is indicated robust property.

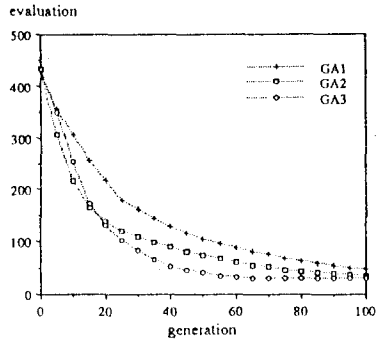

Fig.4 Convergence of the evaluation function

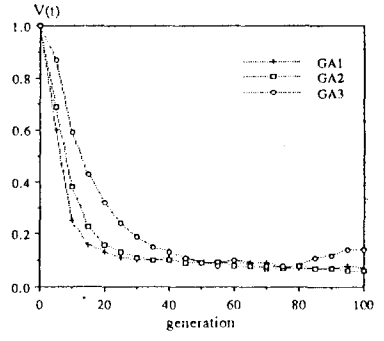

Fig.5 Variations of the degree of diversity $V(t)$

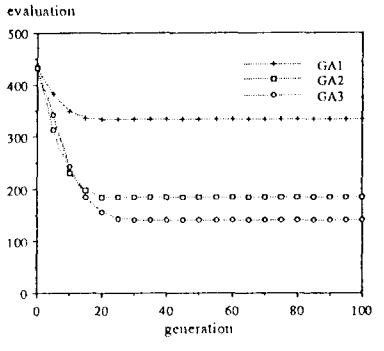

Fig.6 Convergence of the evaluation function without mutation

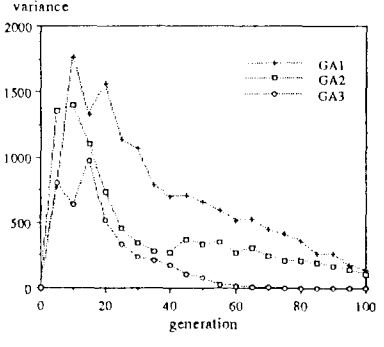

Fig.7 Variances of the evaluation function

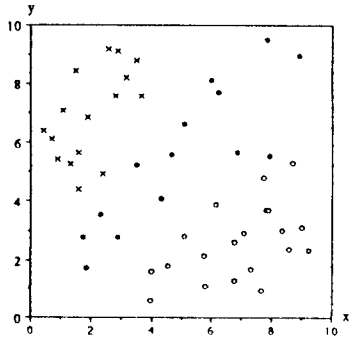

Fig.8 Classification result of GA3

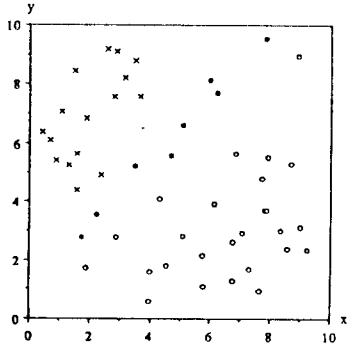

Fig.9 Classification result of the MST method

5. Conclusion

In this paper, we have proposed the combinatorial optimization method using GA for the clustering, especially for the clustering not based on the Euclidean distance. Moreover, in order to improve a search performance of GA, we have introduced the concept of the degree of population diversity, and we have used this index for control parameters of the genetic operators and proposed a method of calculation for the fitness of an individual.

Numerical simulations have shown the followings: 1) The GA using the concept of the degree of population diversity is superior to the traditional GA with respect to the convergence of the evaluation function and the reduction of the variance of the evaluation function; 2) When data have observation noises, the GA using the proposed method shows robustness for data classification. From these results, we conclude that GA using the degree of population diversity is useful for the clustering not based on the Euclidean distance.

Since the concept of the degree of population diversity is not limited to clustering, our next work will be application of the present method to many problems to investigate its general usefulness.

References

[1] J. H. Holland : *Adaptation in natural and artificial systems*, University of Michigan Press 1975.

[2] D. E. Goldberg : *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley 1989.

[3] L. Davis : *Handbook of Genetic Algorithms*, Von Nostrand Reinhold 1991.

[4] Z. Michalewicz : *Genetic Algorithms + Data Structures = Evolution Programs*, Springer Verlag 1992.

[5] H. C. Romesburg : *Cluster Analysis for Researchers*, Robert E. Krieger Publishing Company, Inc. 1989.

[6] A. K. Jain, and R. C. Dubes : *Algorithms for Clustering Data*, Prentice Hall, Inc. 1988.

[7] R. O. Duda, and P. E. Hart : *Pattern Classification and Scene Analysis*, John Wiley & Sons, Inc. 1973.

[8] M. R. Anderberg : *Cluster Analysis for Applications*, Academic Press, Inc. 1973.

[9] V. V. Raghavan, and K. Birchard : "A Clustering Strategy Based on a Formalism of the Reproductive Process in Natural Systems", in *Proc. of the second International Conf. on Information Retrieval*, pp.10–22 1979.

[10] J. D. Kelly, Jr. and L. Davis : "Hybridizing the Genetic Algorithm and the Nearest Neighbors Classification Algorithm", in *Proc. of the fourth International Conf. on Information Retrieval*, pp.377–383 1991.

[11] J. N. Bhuyan, V. V. Raghavan, and V. K. Elayavalli : "Genetic Algorithm for Clustering with an Ordered Representation", in *Proc. of the fourth International Conf. on Information Retrieval*, pp.408–417 1991.

[12] B. S. Duran, and P. L. Odell : *Cluster Analysis A Survey*, Springer Verlag 1974.