Data Clustring Using A New CGA(Chaotic-Generic Algorithm) Approach

Reza Javanmard Alitappeh
Islamic Azad University Sari Branch, Sari, Iran
rezajavanmard64@Gmail.com

Mohammad Mehdi Ebadzadeh
AmirKabir University, Tehran, Iran
ebadzadeh@aut.ac.ir

Abstract

Clustering is the process of dividing a set of input data into a number of subgroups. The members of each subgroup are similar to each other but different from members of other subgroups. The genetic algorithm has enjoyed many applications in clustering data. One of these applications is the clustering of images. The problem with the earlier methods used in clustering images was in selecting initial clusters. In this article it has been tried to develop a set of populations (i.e., cluster centers) using the clonal selection of artificial immune system, and to obtain the final clustering of clusters and the main image among a large number of clusters through the use the K-means and the K- nearest neighbor algorithms. Moreover, chaotic model has also been used to create diversity both in the original population and in the populations produced through the repetition of generations. The algorithms in the paper have been executed on satellite images; and the implementation results showed that the algorithm works well.

Keywords: *image clustering, genetic algorithm, chaotic model, Satellite image, Chaos-Genetic Algorithm, Fitness.*

1. Introduction

The task of grouping data into similar clusters is a kind of unsupervised learning and has applications in many areas. The technique presented in this article makes use of information obtained from pixels to process images in the machine vision application, for example. As it is known, various methods have been proposed for clustering data, especially data related to images. Clustering images has a wide range of applications, such as clustering medical images, image concerning nature like distinguishing forests and mountains, satellite images, etc. Specifically, in the area of satellite images, today important applications are used such as weather forecasting, studying different layers of the surface of the earth by geologists in an effort to discover aquifers, oil, gas, or to study volcanoes. In this article, the use of genetic algorithms in clustering images will be discussed. Using this technique, the initial population is produced in a standard random way; and chaotic model was used for this purpose.

GA (Genetic Algorithm) has advantages such as flexibility, adaptability, and stability in solving optimization problems which are very difficult to solve using conventional optimization methods. Despite all these advantages, GA has two shortcomings. Firstly, it tends to converge prematurely and secondly, it performs much iteration to reach the optimum goal. To overcome these problems related to searching, a composite genetic-chaotic model was used in this research. Another merit of the proposed algorithm is that a large number of initial points are used (as centers of initial clusters). Clonal selection

was the inspiration to do this. In the proposed algorithm selections were always clonal based and the best of every generation were repeated in each clone. The advantage of doing this is that in the end we would have a bigger set to choose our optimal answer from. This is because of the K-Means algorithm usage in our final clustering. The remaining parts of the article are organized as follows: in section two, the problems of clustering was addressed. In section three the genetic algorithm was proposed. Then, chaotic model was analyzed in section four and the composite chaotic-genetic algorithm was presented in section five. Section six discussed the proposed composite genetic-chaotic algorithm in clustering images in detail based on its implementation. The experimental results were presented in section seven. Finally, conclusions were included in section eight.

2. Clustering Problem

The diversity of clustering applications has resulted in tackling different problems. The purpose of all developed algorithms is to divide a set of data into subgroups. The members of each subgroup are similar to each other but different from members in other subgroups. there are many different methods for calculating similarities and differences which poses many different problems. In proposed algorithm, the clustering problem was considered as the process of dividing a data set into a number of subgroups so that the result equation is minimized. Generally, the equation used was as follows (although other parameters have been employed as well, which will be explained in section 3-b-4)[1]:

$$E = \sum_{k=1}^{K} \sum_{X \in CK} \sum_{a=1}^{A} (X_a - mK_a)^2 \qquad (1)$$

Where K is the number of clusters ,X the data set considered ,CK the kth cluster indicator, mK the average of the kth cluster ,and A the total number of features in a data set. The above equation was used to find the Euclidean distance from the center of each data set to the center of another data set. Minimizing the general distance of a cluster leads to the optimum answer.

The process of finding this class is known as NP-Hard1. In this article there is a discussion on the images received from satellites with the purpose of separating natural land surfaces for geological and geographical uses with the aim of separating man-made land surfaces for urban uses,etc.

3. The Genetic Algorithm

Evolutionary Calculations (EC) are an area in computer sciences in which biological processes are employed as a model for solving problems through the use of computers [2].

The genetic algorithm (GA), which was first proposed by John Holland in 1960, is a class of EC which uses the concept of evolution. The main application of GA is in finding a balance between exploration and extraction of an optimum solution in the search space related to the problem. The general procedure used in the genetic algorithm is as follows:

First, taking the space of the problem into consideration, an initial population of chromosomes (containing genes) is built up. Each chromosome is a solution which the GA tries to optimize. The GA carries out the exploration/extraction processes in the

search space of the problem while the evolution of the population is taking place during the succeeding generations.

During these generations, new populations of chromosomes are produced by the mutation and recombination operators and among them, the best are chosen to replace their parents [4]. This process continues till convergence is achieved. Figure 1 shows the general schema of this process. One of the important stages in GA is the selection of those left to be used for the next generation. This is carried out by taking into account the quality criterion of the chromosome, which is called the fitness of the chromosome.

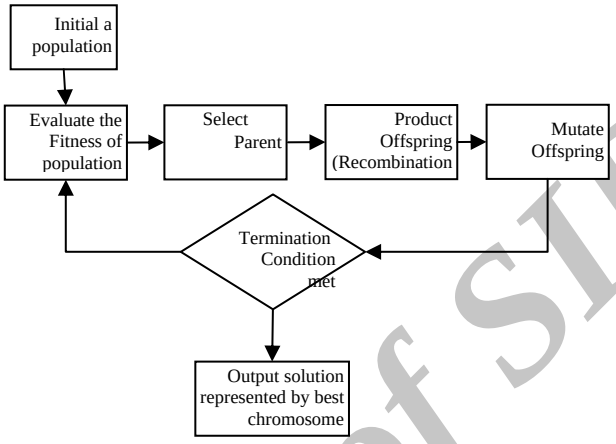

Figure 1. *Process of Genetic Algorithm trend*

4. The Caotic Model

The chaos theory owes its development mostly to the studies of Henry Poincare, Edward Lorenz, Benoa Mendelbroot, and Michael Fagenbaum. Another branch of the chaos theory which is used in quantum physics is called the quantum chaos. Chaos literally means irregularity and disturbance and has a negative connotation. However, with the development of a new viewpoint called the chaos theory, and with the elucidation of its scientific and theoretical dimensions, today this theory refers to the unpredictable and random aspects of dynamic phenomena [5].

In this article, a new and improved strategy was introduced. This strategy was a new genetic algorithm containing the chaos variable, used the random internal feature of chaotic repetitions to obtain early local optimums and also to increase the speed of the genetic algorithm [6]. Moreover, the chaos-genetic algorithm (CGA) combined the chaotic-like local search with the non- dominant ordering genetic algorithms to solve multi-purpose optimization problems [7]. In fact, the CGA is a new composite strategy for overcoming the weaknesses of the simple genetic algorithm (SGA) [8]. In the article chaos is used in initialization to improve the quality of the components and to create population diversity.

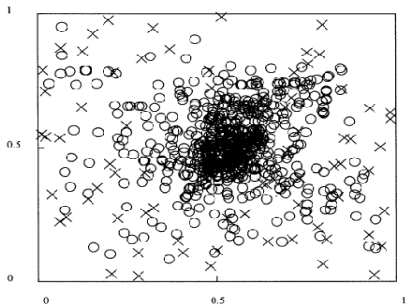

Figure 2. *Distribution of population during standard GA process.X First generation population O Next generation population*

In Figure 2, it can be seen that the distribution of sub-generations of members slowly concentrates towards the center of the defined space.

4.1 Chaos Mapping Operator

Assume that the independent variable vector shown by X contains X elements. These elements, which are the function vectors of X, are called function parameters and are represented as $X_1,X_2\ldots,X_n$. Therefore, a minimal search problem can be expressed as follows:

$$\min_x f(x_1, x_2, \ldots x_n) \qquad (2)$$
$$x_i \in (a_i, b_i) i = 1,2,\ldots,n$$

The F function is a relationship value between the X-dependent variables and the dependent variable needs optimization.

A space n was defined as so having a range of $\{(a_i\ b_i), i=1,2,\ldots,n\}$,where $[a_1,a_2,\ldots,a_n]^t$ is the lower limit shown by the vector a, and $[b_1,b_2,\ldots b_n]^t$ is the upper limit shown by the vector b.

The evolutionary process of chaos variables would be defined by an extension equation like the following:

$$cx_i^{k+1} = 4cx_i^{(k)} + (1 - cx_i^{(k)}), \qquad (3)$$
$$i=1,2,\ldots,n$$

Where cx_i is the ith chaos variable and both K and K+1 show the number of iterations.

The defined variable by the above equation is the chaos variable, and cx_i are distributed in the interval [0, 1], provided that $(0, 1) \in C\ x_i^{(t)}$ and $(0.25,0.5,0.75) \notin cx_i^{(t)}$ so as to ensure that the evolutionary process is carried out properly. The following process was performed in order to make use of the advantages of chaos in evolution:

1) The working parameters $x_i^{(1)}$,i=1,2,…n, were a linear mapping of the defined space S_0 into a normalized chaos space that was defined as $\{(0,1), i=1,2,\ldots n\}$ and acted as the operator of the linear mappings.

$$cx_i = \frac{1}{b_i - a_i}(x_i - a_i), \qquad (4)$$
$$i = 1,2,\ldots,n$$

Note that the first iteration was $x_i^{(1)}$ and the result of this iteration would be the first iteration for the variable $cx_i^{(1)}$.

2) The next iteration of the chaos variable ($cx_i^{(2)}$) was obtained by applying the iteration operator defined in Eq. 5 on the variables $cx_i^{(1)}$.

3) The chaos variables $cx_i^{(2)}$ were linearly mapped from the space cs_1 into the space S_0 by the backward mapping operator in order to produce the working parameter x_i through the use of the following equation:

$$x_i^{(2)} = a_i + cx_i^{(2)}(b_i - a_i), i = 1,2, \dots, n. \qquad (5)$$

Therefore, the working vector of the kth iteration (x_k) was defined by steps 3, 4, and 5 for all workings parameters: the working vector x_k was mapped forward and backward in a chaotic mapping so that the (k+1)th iteration, that is the working vector x_{k+1} ,was obtained. This whole process is called the chaos mapping operator and it is represented by Ψ. The flowchart of the chaos mapping operator (CMO) was shown in Fig 3.

$$X^{(K+1)} = \psi(X^{(K)}) \qquad (6)$$

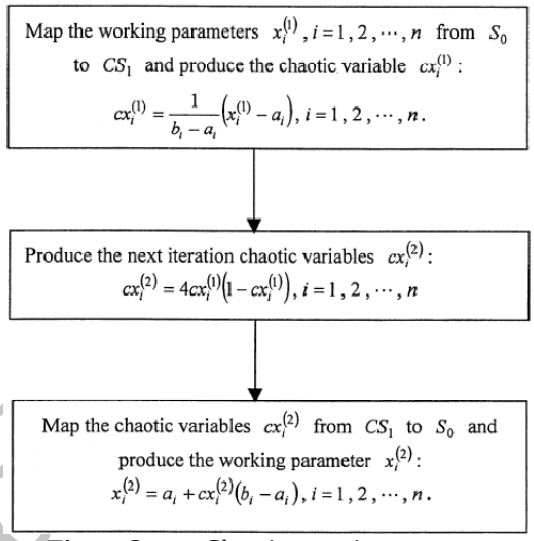

Figure 3. *Chaotic mapping operator*

5. Chaos-Genetic Algorithm

The execution of the CGA would be described as follows:

1) Determine the number of the population np the crossover probability Pc , the mutation probability Pm, and the maximum number of generation tm. Set the current number of generation t= 1, and the first generation of individuals $x_1^{(t)}$, $x_2^{(t)}$, . . ., $x_{n_p}^{(t)}$,which are randomly produced with the values in the space S0, and are denoted by $X_t^{(1)} = \{x_1^{(t)}, x_2^{(t)}, \dots, x_{n_p}^{(t)}\}$.

2) The chaotic individuals represented by $X_t^{(2)} = \{x_1(t)^{(2)}, x_2(t)^{(2)}, \dots, x_{n_p}(t)\}$are produced by CMO (ψ) for all the individuals of $X_t^{(1)}$, i.e.

$X_i(t)^{(2)} = \psi(x_i(t))$, i= 1, 2, . . .,$n_p$.

3) Randomly select n_p individuals from the set consisting of $X_t^{(1)}$ and $X_t^{(2)}$ to form a reproduction set $X_t^{(r)}$ using the genetic operation of proportionate reproduction to meet the well-known Darwinian theory of survival of the fittest.

4) Produce the new population $X_t^{(3)}$ from $X_t^{(r)}$ through the genetic crossover operation .

5) Produce the new population $X_t^{(4)}$ from $X_t^{(3)}$ through the mutation operation.

6) Set t=t+1 produced, and $X_t^{(1)}= X_{t-1}^{(4)}$ is executed and repeat steps 2- 5 as long as the number of generations is smaller than the allowable maximum number tm. . The related flowchart was shown in Figure. 4. In Figure. 5, the distribution of members of the sub-generation produced by the CGA was explained.

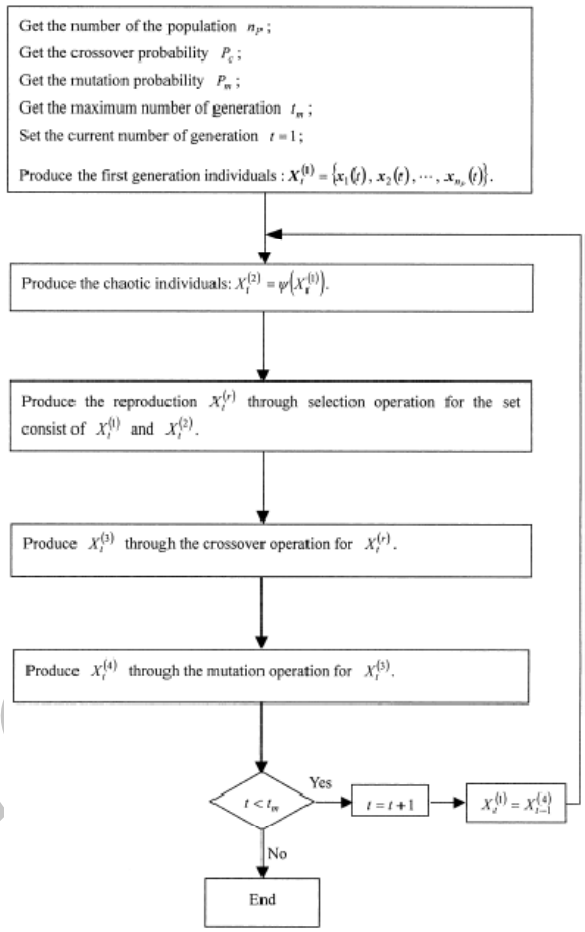

Figure 4. *Chaoc-Genetic algorithm*

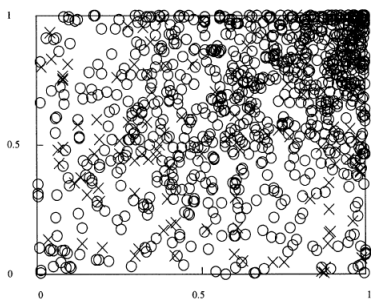

Figure 5. *Distribution of population via Chaos-Genetic algorithm X First generation population O Next generation population*

It appears from Figure. 5 that members of sub-generations were distributed almost throughout the whole defined space and that they didn't have concentration towards any center [9].

Figure 6. *A satellite image of the size 128*128 and with 32 frequency bands*

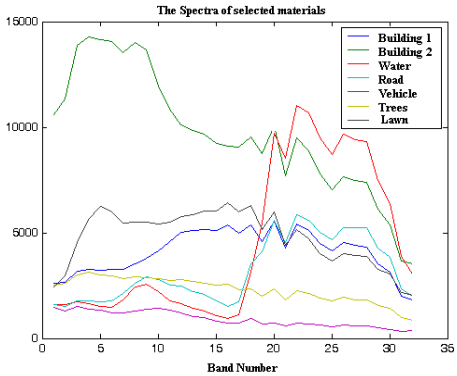

Figure 7. *The 32 band of the seven centers of the clusters*

6. Proposed Chaos-Genetic Algorithm In Image Clustering

6.1 clustering satellite image

As was explained in previous section, clustering satellite images has many applications in various fields such as geography, geology, meteorology, etc. We have conducted our experiments on a satellite image of the size 128*128 and with 32 frequency bands (Figure. 6). In other words, for each pixel of the image, there was a vector having 32 features. This image, compared with images used for clustering color features in other fields, involved more calculations. An example of such images, which has seven centers and was used for the purpose of distinguishing water, roofs, roads, trees, vehicles and grass from each other, was shown in Figure 6. The 32 features relating to the seven pixels of the center of the clusters were shown in Figure 7. As can be seen in the figures, the degree of overlapping of the lines needs reduction.

6.2 Proposed algorithm

The used algorithm was as follows: first, the initial population to start the evolution of the population was used based on fitness, only the fittest chromosomes were left. In the next stage, among the 100 remained genes in chromosomes, 7 centers were extracted based on k-means method. Finally, classification was done according to the

degree of distance of each pixel in the image, on the original image. It must be noted that last phase was performed by default using the squares of Euclidean distances. It must be mentioned that at the stage while calculation of the fitness of chromosomes, due to the fact that a long time was needed to examine all the pixels of the image, 700 pixels were used, which were selected randomly, for this purpose, and thereby the time needed to do the calculation was reduced. Steps in proposed algorithm can be divided in 6 steps as follows:

1) Representation

One of the first stages of evolutionary algorithm development is re-representation or, in other words, the data modeling of the space of the problem. In this paper, chromosomes were considered having the length of 100 which means each chromosome had 100 centers. Each gene had two dimensions which were indicators of the positional coordinates of the pixels (Figure. 8). The x_i and y_i 32-dimension of feature vector was recovered from the main file, as was explained in the last section.

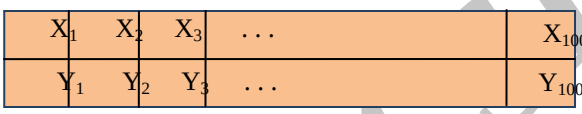

Figure 8. *Representaton of chromosom that used in this article*

2) The initial population

From the initial population, 1000-gene chromosome was considered in which each gene had two points x_i and y_i. An additional gene was added for the maintenance of the fitness of the chromosome. The initial population was consisted of 50 chromosomes. The genes had values varying from 1 to 127. After building the initial population, it was taken to the chaos space using the chaos mapping operator, where was changed to the chaos model. Subsequently, this model was returned to the original space. This operation was performed once in each iteration.

3) The mutation and recombination operators

To perform the recombination, because of the re-representation used, a new method was employed: the chromosome was divided by determining four random points and were substituted by each other, as in the standard method defined for the GA in reproducing new children, we replaced them with each other, with row replacement as well as column replacement. In other words, in the new child, the positions of the x and y parents may have changed. As can be seen in Figure 9, to produce a new child, the feature of producing new and random children was considered.

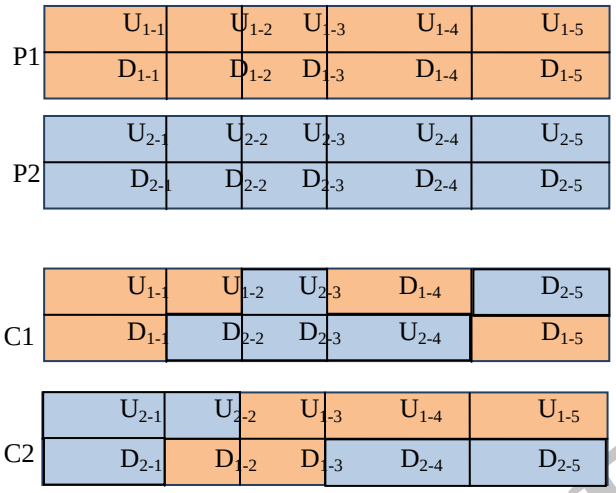

Figure 9. *Recombination operator (P1,P2: Parents, C1,C2: Childs)*

The mutation operator acted on one chromosome and produced one chromosome. The following method was used for this purpose; a part of the x was taken and turned in a rotational mutation.

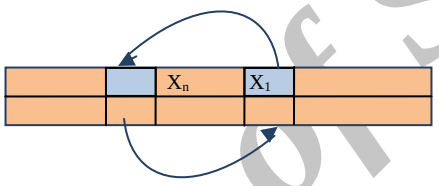

Figure 10. *Mutation operator*

4) Evaluation of fitness

The fitness function used is as follows:

$$J=\sum_{i=1}^{128}\sum_{j=1}^{128}Max_{k=1}^{100}Cos((Xij,Vk))-a\sum_{i=1}^{128}\sum_{j=1}^{128}Min_{k=1}^{100}\left\|Xij-Vk\right\| \quad (8)$$

Equation 8 showed that, on the left side the cosine distances of all the genes in the chromosomes was first calculated and the maximum distance among them was chosen. In other words, for each x and y in the chromosome, distances from all the pixels in the image were calculated. On the right side, the Euclidean distances of all the genes from a chromosome are calculated and the genes having the minimum distance was chosen. In other words, the maximum distance of each point in the image (as stated earlier, only 700 random points were chosen as representatives of the image) from the centers was first calculated (the seven values for each chromosome), and then added all up. To calculate the distance or, in other words, the similarity of each point of the center, the equations for cosine distances was used which was defined as follows:

$X_{i,j}$ refered to the 32-dimension eigenvector of a point in the image ,and V_k refered to the eigenvector of the 100 cluster centers present in the chromosome. Cos (X_{ij},Y_k) was equal to the internal product of the two 32-dimension eigenvectors of V_k and X_{ij}.

5) Offspring selection

The elitist method was used among various methods for choosing the survivor. This method, the decisions are made on the basis of the fitness of populations of course, other methods were applied in our experiments as well.

6) Using of k-means clustering

After optimizing a population of cluster centers by using evolution, in the next stage clustered these centers were clustered by using the k-means algorithm to obtain the final centers [3]. Finally, the obtained clusterswere used to get the final image. the image was clustered using the k-nearest neighbor algorithm.

7) Experimental Result

As was explained in previous sections, a satellite image with 32 features in each pixel was used in our experiments. Furthermore, the following parameters were used in implementing the proposed algorithm. (Table 1)

Figure 11. *Parameters setting*

parameter	value
A (used in Eq. 8)	0.61
Max Generation	15
Population size	20
P_m (mutation rate)	0.8
P_c (recombination rate)	0.2
γ (max generated child)	7γ
μ (survivor number)	5

The clustered and main images were showen in Figure. 11. Obviously, in subjective way, clustered image based on proposed algorithm was very close to goal image.

32-Dim of clustered image and evaluated fitness during genetic algorithm process were showen in Figure 12, 13. Distribution of population in special generation and fitness of best chromosome that produced 7 final centers were presented in Figure 14 and 15. As showed in Figure. 14 a good distribution was occurred during proposed CGA.

8. Conclusions and future research

As was observed in the above experiments, the time needed to do the calculations increased due to the usage of a cosine and a Euclidean function. However, good answers were obtained in the outputs.

Broadly speaking, since a set of centers were developed and at the end the best was chosen, the method in this paper was proved to be a good one. There appears to be other methods for the final division of the centers to seven main centers. As a topic for further research, the chaos mutation operator can be used instead of standard mutation. This will certainly lead to finding better answers.

Reference

[1] Q. Ding, J. Gasvoda, "A Genetic Algorithm for Clustering on Image Data", International Journal of Computational Intelligence Winter 2005

[2] K. Venkatalakshmi, P. Anisha Prasy, R. Maragathavalli and A. MercyShalinie, "Multispectral Image Clustering Using Enhanced Genetic K-Means Algorithm", Information Technology Journal 2007 Asian Network for Scientific Information

[3] V. Ramos, F. Muge "Image Colour Segmentation by Genetic Algorithms ", Accepted in RecPad'2000 - Proc. of the 11th Portuguese Conference on Pattern Recognition

[4] T. Back, D. B Fogel and Z. Michalewicz" Evolutionary Computation Advanced Algorithms and Operators" 2000, ISBN 0750306653

[5] H. J. Korsch , H.J Jodl. Chaos A Program Collection for the PC.Library of Congress Control Number: 2007940051.ISBN 978-3-540-74866-3 Springer Berlin eidelberg New York 2008

[6] C. T. Cheng, W.C. Wang, D.M. Xu K. W." Chau. Optimizing Hydropower Reservoir Operation Using Hybrid Genetic Algorithm and Chaos. Water Resour Manage (2007) 22:895–909

[7] R. Qi, F. Qian, S. Li, Z. Wang. " Chaos-Genetic nAlgorithm for Multiobjective Optimization.", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China

[8] Y. Yong, S. Wanxing, W. Sunan ." Study of chaos genetic algorithms and its application in neural networks." Xi'an Jiaotong Univ., China. Proceedings. 2002 IEEE Region

[9] X. F. Yan, D. Z. Chen, S. X. Hu," Chaos-genetic algorithms for optimizing the operating conditions based on RBF-PLS.", model Computers and Chemical Engineering 27 (2003) 1393-1404

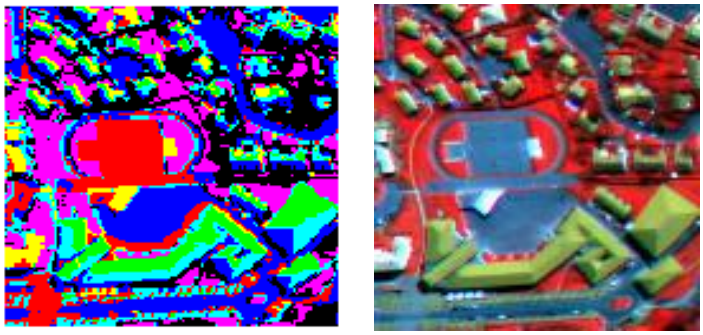

Figure 12. *Right: clustered image (goal), Left: clustered image based on proposed GCA*

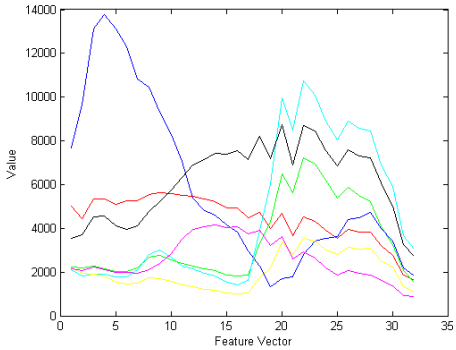

Figure 13. *The 32 band of the seven centers of the clusters*

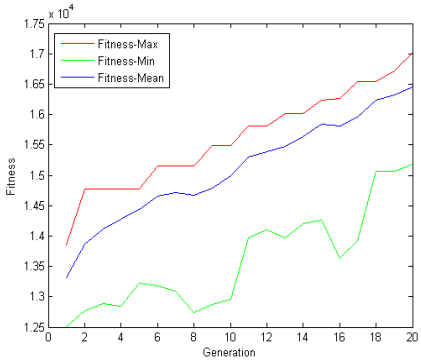

Figure 14. *Fitness during GA iterations*

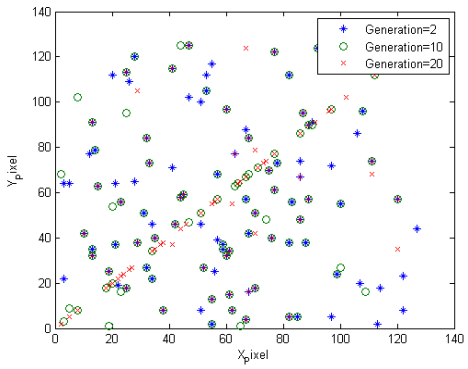

Figure 15. *Distribution of population during generation 2,10 and 20 based on Chaoc model*

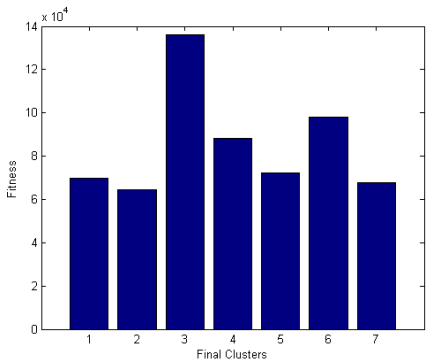

Figure 16. *The fitness of centers obtained from best chromosome*