Rep. Fac. Sci. Engrg.
Saga Univ.
36-1 (2007),17-23

Reports of the Faculty of Science and Engineering,
Saga University, Vol. 36, No.1, 2007

ISODATA clustering with parameter (threshold for merge and split) estimation based on GA: Genetic Algorithm

By
Kohei Arai* and XianQiang Bu **

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. It is found that the experimental results show that clustering error of the proposed method is 2 to 10 times much smaller than that of the existing method. It is also found that the elite selection strategy is superior to the average selection through experiments.

Key words: ISODATA clustering, Genetic Algorithm: GA,

1. 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.

2. Proposed method

Received on Apr.28 2007
*Department of Information Science
**Graduate School of Science and Engineering

2.1 K-means

The k-mean method is that of the non-hierarchical type clustering method proposed by MacQueen, Anderberg [14], Forgy and others in the 1960s [15]. Based on the given initial cluster center of gravity, this method uses the average of a cluster and classifies [which were given / k] it in the number of clusters. The 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. Actually Choosing much different initial value, clustering and choosing a good thing from among the obtained results is often performed.

2.2 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 excepts 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 center 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 1st 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.

2.3 Evaluation of clustering result

A different clustering result is obtained from the separate clustering method in many cases. Also by the same clustering 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. A pseudo F value (pseudo F statistic) is one of the valuation bases often used. A pseudo F value is defined by the following equation.

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

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. A pseudo F value 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 the value of F is large, a clustering result shows a good thing. However, since cluster distribution assumes it as the convex function in the pseudo F value, 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.

$$E = \frac{n-c}{n} \times 100\% \quad (2)$$

2.3 Heredity algorithms

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 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. Adaptive value (fitness) is searched for based on the adaptive value function beforehand defined to each individual.
3. Selection is performed based on adaptive 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 adaptive value of an individual is high, but the low individual of adaptive value may also survive the next generation. This is a role which controls lapsing into a partial solution. It has achieved.
   - Tournament selection: It is the method of repeating this process until it selects a certain number of individuals at random from the population, adaptive value chooses the best thing in it and the population's number of individuals is obtained.
   - Elite strategy : How many individuals with the maximum adaptive 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. Adaptive 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, is excellent in global searching ability, and is widely applied to various optimization or a search problem.

2.4 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) 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, single modal normal distribution crossing method (Unimodal Normal Distribution crossover: UNDX):, 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 equation.

$$A = \min(a,b) - \alpha|a-b|$$
$$B = \max(a,b) + \alpha|a-b| \quad (3)$$

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

- Mutation by a normal distribution: Happen 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 an 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.

2.5 The proposed 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 adaptive 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 [in ISODATA], it decided to avoid an above-mentioned problem by controlling this threshold. The convectiveness 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.

2.5.1 Partial mean distance

As partial mean distance is shown in the lower portion of 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 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 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.

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.

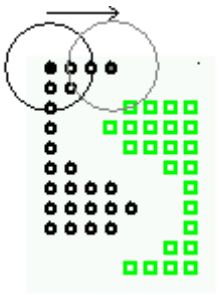

Fig. 1Partial mean distance

2.5.2 The 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, 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 proposed 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 proposed 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 RCGA by the proposed 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.

2.5.3 Selection of adaptive value function

Since the cluster that a pseudo F value 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 adaptive 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 / the sum of partial mean distance all over the districts] -- since only the crevice within the limits between parts will be observed if an adaptive value function setup is carried out, a cluster may not become a lump.

$$\text{Fitness} = s + \frac{m}{d} \tag{4}$$

In this, s expresses the sum of partial similarity all over the districts, d expresses the sum of partial mean distance all over the districts, and m expresses weight. And s is calculated by the following equation

$$s = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \delta_{i,j} (2r - \|l_i - l_j\|)$$
$$\delta_{i,j} = \begin{cases} 1 & C_i = C_j \\ 0 & C_i \neq C_j \end{cases} \tag{5}$$

d is calculated by the following equation

$$d = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \delta_{i,j} \|l_i - l_j\|$$
$$\delta_{i,j} = \begin{cases} 1 & C_i = C_j \\ \frac{1}{2} & C_i \neq C_j \end{cases} \tag{6}$$

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 proposed 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.

2.5.4 A setup of the parameter of 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%.

End conditions, the elite, five-generation maintaining t as a thing and five generations of differences of the average adaptive value's and the elite's adaptive 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.

3. Experiments

3.1 Experiment 1

The proposed method experiments by making the data of simple convectiveness 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 the lower portion of Fig. 2, it experiments using the data containing two clusters of convectiveness. It clusters by the ISODATA method and the proposed method with a random parameter, and the result of clustering is compared.

Fig. 2 Former data

The result of an experiment is shown like a lower portion of figure. The error which cannot understand the cluster of concaveness 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 proposed method, it turns out that an error becomes 0% and it can classify according to division and fusion correctly with a curve.

Fig..3 The ISODATA method (Random)

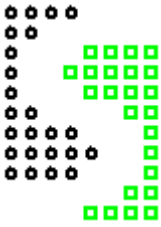

Fig.4 The proposed method

3.2 Experiment 2

Next, it experiments using the Iris data set of UCI, 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 adaptive value of 50 generations, i.e., the process of convergence, is shown in Fig. 5.

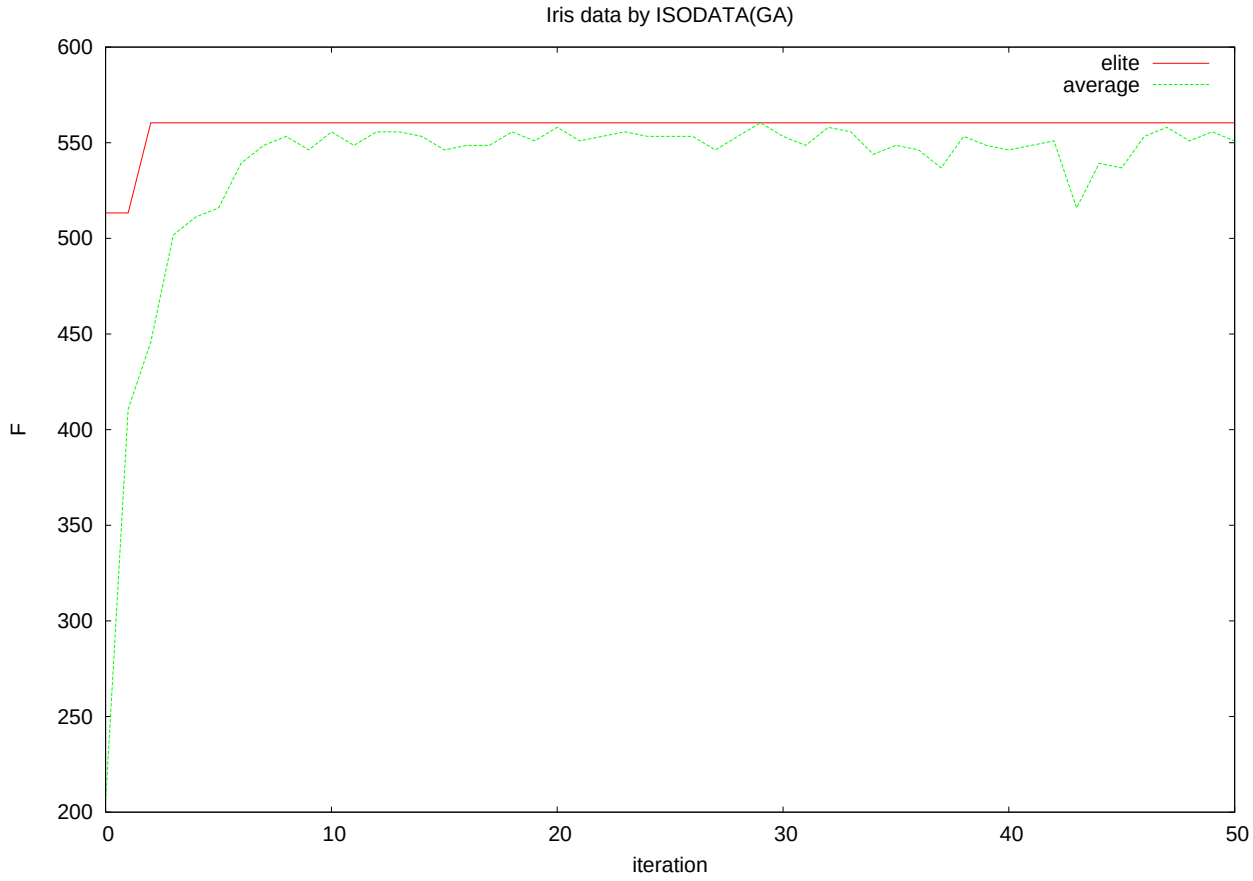

Fig.5 Convergence process of GA

In this figure, a red line expresses the elite's adaptive value, and expresses the average of adaptive value with the green line. This figure shows being completed by the average of adaptive 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 a lower table.

Table 1 When number [of categories]-in agreement with number of target clusters

Error (%)	The ISODATA method (Random)	The proposed method
Iris	21.33	10.67
Wine	34.65	2.81
Ruspini	12.43	0.00
New thyroid	31.63	11.16

In Table 1, compared with the conventional ISODATA method, the direction of the proposed method shows that the error decreases 18.85%, respectively. Although optimizing took about 10000-time long time from the conventional ISODATA method, accuracy went up certainly.

4. Concluding remarks

In this research, it was mentioned by the experiment that the proposed method is predominant to the conventional method. Although computation time became long by the proposed method, the result of an experiment showed that the better result of the degree of separation between clusters could be obtained. By the proposed method, it also turned out that it can respond also when cluster distribution is a concave function. This

thinks that it originates in being optimized by GA also not only in the initial cluster center of ISODATA but in merge and split of clusters. It is found that the experimental results show that clustering error of the proposed method is 2 to 10 times much smaller than that of the existing method. It is also found that the elite selection strategy is superior to the average selection through experiments.

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 in Neural Information Processing Systems 7, Cambridge, MA: The MIT Press, 1995

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

[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] 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.

[15] Forgy, E.W., Cluster analysis of multivariate data: Efficiency versus interpretability," Biometric Society Meetings, Riverside, CA. Abstract in Biomatrics, 21, 768, 1965.