J. Appl. Environ. Biol. Sci., 6(1)38-45, 2016

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

An Improved Grouping Genetic Algorithm

S.A. Shahzadeh Fazeli* and S. Barkhordari

Parallel Processing Laboratory, Faculty of Mathematics, Yazd University, Yazd, Iran

Received: September 23, 2015
Accepted: November 15, 2015

ABSTRACT

In this paper we introduce and explore some clustering algorithms, an algorithm that can provide acceptable results. The merits of the algorithm we consider is its fast performance, by using the appropriate feature of the previous methods and dispel some of the defects of classical Genetic Algorithm for grouped issues.
KEYWORDS: Clustering algorithms, Genetic algorithm, Grouping Genetic Algorithm, DBSCAN.

1.INTRODUCTION

The clustering is one of the main data mining methods; yet there is not a suitable method to replace it. In general, the data objects are grouped into distinct groups (clusters), this is called clustering. Clustering is used in various fields such as pattern recognition, engineering, image processing, etc.

Today, due to the large amount of information, clustering in data mining has a special place. That is why finding the right algorithm in this area is very valuable. Genetic Algorithms (GAs) are examples of algorithms that are used in this field and have improved tremendously in the past two decades.

In the past years, several clustering algorithms based on GAs have been developed. These algorithms use different representations for the clustering solutions. One kind of the methods uses a straight forward encoding, in which the chromosome is encoded as a string of length *n,* where *n* is the number of data points, and the element of the chromosome denotes the cluster number that data point belongs to, such as used in [1, 2]. This approach does not reduce the size of the search space and searching the optimal solution can be onerous when the data points profile rate. It is for this reason that some researchers use a relatively indirect approach where the chromosome encodes the centers of the clusters, and each data is subsequently assigned to the closest cluster center [3–8]. Tseng and Yang [9] proposes GAs for the clustering problem that is suitable for clustering the data with compact spherical clusters. This algorithm can automatically evolve the number of clusters. But the number of clusters obtained is influenced by some parameters used by the algorithm. In [10, 11], a hyper-quad tree is employed to denote a set of centers. The weakness of this approach is the need to establish the sets of centers occupying the same region of space that can be very time-consuming and prone to bad solutions when such sets are inappropriately selected. Yet, other researchers are seeking to the hybridization between GAs and K-means with interest in using GAs to feature selection for clustering. In [12, 13], GAs are used to evolve the features serving as the input for the K-means algorithm. The clustering solutions of K-means algorithm are then evaluated and the resulting objective function on values is fed back to the GAs. Two co-evolutionary algorithms are used for the feature weighting in [14].

In [15] an evolutionary algorithm for clustering is described, where the objective function is constructed through a message-based similarity function, which simulates messaging between objects and optimal centroids of the clusters. The performance of the approach is shown in different synthetic examples and also in real data sets from UCI repository [16].

GA is inspired by Darwin's theory of evolution; it is based on the ultimate survival of natural selection getting started with an initial population of solutions and using the selection operators, crossover and mutation which will try to improve them. The Genetic Algorithm consists of the following steps, [17, 18]:

Step1. To create initial population, a group of chromosomes is randomly generated.
Step2. Suitability of each population is calculated.
Step3. According to the suitability, two people are selected as a parent.
Step4. Parental chromosomes are combined to create new children.
Step5. According to the probability assigned to the mutation, the children are mutated.
Step6. The new population is replaced with current population.
Step7. If the conditions are provided the algorithm stops, otherwise returns to step3,

*Corresponding author: S.A. Shahzadeh Fazeli, Parallel processing Laboratory, Faculty of Mathematics, Yazd University, Yazd, Iran fazeli@yazd.ac.ir

2. Classical Genetic Algorithm

In the following the Classical Genetic Algorithm for clustering issues will be discussed.

2.1 Classical Genetic Algorithm for clustering problem

Genetic Algorithms have been created with some changes in encoding Genetic Algorithm to be suitable for classification issues. The groups are shown by genes on the chromosomes. For example, if the chromosome is BCAC, the first gene is B, the second gene C, the third gene A and gene C is fourth. This chromosome indicates that the first data is in Group B, the second in Group C, the third in Group A and fourth is in Group C.

The purpose of Genetic Algorithms is creation of children with better fitness than their parents. One of the simplest and classical crossover operator used is a single point crossover. To demonstrate the performance of this operator, consider two chromosomes ACBB and BCAA. Single point crossover randomly selects, which is called cutting and switching between sectors is performed as follows:

The first parent: ACB |B first child: ACBA

Second parent: BCA |A second child: BCA B

Suppose first child is selected for mutation. Mutation operator selects and changes the amount of second gene randomly, as follows:

ACBA→ ABBA

2.2 The weaknesses of Classical Genetic Algorithm for clustering problem

Emmanuel Falkenauer, [19, 20], has propounded defects caused by acts of classical Genetic Algorithm on grouped issues. In his speech, this encoding method can have very adverse effect on the data. It is also possible that the solution of the crossover operator does not have even one attribute from both parents. For example, when parents selected for single-point crossover solutions offer the same solution to a classification problem, the operator may create children that are not valid solutions. Similarly, when a Genetic Algorithm is reached to the perfect solution of the problem, standard mutation with changes in successful chromosome, reduces the severity of merit solutions and possibly eliminate it from population.

3. Grouping Genetic Algorithm

Grouping Genetic Algorithm (GGA) was created in 1992 by Falkenauer [19], due to problems in Classical Genetic Algorithm for clustering issues. Grouping Genetic Algorithm corrects model of display and mutation and crossover operators with new encoding to produce solutions with high quality for every grouping problem.

3.1 Encoding Grouping genetic algorithms

Falkenauer's encoding of Genetic Algorithm includes a new section that is called grouping added to the original chromosome. For example, chromosome CACB, ABC groups are added such that any combination of the ABC that includes all groups is valid. In this encoding, the chromosomes can have a variable length. Chromosomes use several distinct types of elements in various combinations of groups with all sizes. Adding group section is the main difference of this encoding with previous methods, but in addition the operators of Grouping Genetic Algorithm also change the classification of Genetic Algorithms and run on both section of chromosomes [21].

3.2 The operator of Grouping Genetic Algorithms

Grouping genetic operators perform crossover and mutation with some changes. The standard crossover changes some portions of chromosomes of parents to make their children. When standard crossover is applied in grouping problems, born chromosome can contain one or more groups that share the same content. Since each proposition may belong to a group, this implementation results in invalid chromosomes.

To prevent the creation of invalid children, Falkenauer introduces a new crossover operator. The operator initially selected parents and a point as crossover point in their group section randomly, then the elements belonging to the selected groups was copied from the point of crossover of the first parent to child. Next, consider the second parent. Selected groups of elements belonging to the second parent are copied for children if the first parent is not specified.

When standard encoding used, classical mutation can lead to some problems. If a minor transfers to a group randomly that does not have any similarity with other members, quality solutions can be very effective. Falkenauer mutation in the proposed model is as follows:

A new group created and current group is deleted then some propositions that were selected from their respective groups randomly are combined. For each choice, we should determine the innovative solution to be used. In During repeated generations, repeat the same algorithm.

4. Grouping Genetic Algorithm for clustering problem

In this part, the Falkenauer's proposed algorithm is used for clustering problem. In fact, with particular attention to encoding, genetic operators and the implementation of a local search improve the quality of solutions and finally with an evolutionary island model the parallel algorithms will be discussed.

The general process algorithm is as follows [22]:

Step 1. Form the initial population of chromosomes. The encoding chromosomes must be valid for clustering.
Step 2. Calculate the fitness of each chromosome.
Step 3. Select the parent chromosomes by the operator based on rank [23], are from the population.
Step 4. Perform crossover operation to drive children chromosomes by combining parent chromosomes.
Step 6. Perform mutation to avoid uniformity population, with dynamic probability.
Step 5. Replace the new population that has been created by three former steps,

How to encode: In this algorithm, each individual is shown with two parts:

C = [I | g] where I represents the element and g represents the group section of chromosome. Let X be the input vector:

$$X = [x_1, x_2, x_3, x_4, x_5]$$

Consider the following chromosomes:

[1 3 2 1 4|1234]

Regarding clustering problems, there are four genes in group section. These genes represent a cluster. In the element section, each gene indicates a cluster that is related to it. For example, the first gene in the elements of chromosome is one. This means that x_1 belongs to one cluster. The second gene is three, which means that x_2 belong to the three clusters. This approach is continued to the end similarly [24].

Selection operator: For parental choice, the selection based on the rating is used in accordance with James et al. [25]. Firstly, the quality of chromosomes based on distance functions and indices are calculated. Then according to given score to people, they have to be arranged in descending order. Ri, $i = 1, \dots .\varepsilon$ displays rank of each person on the list and ε is the population number.

Fitness for an individual is defined [24] as follows :

$$F_i = \frac{2 * R_i}{\varepsilon * (\varepsilon + 1)}$$

Crossover operator: The crossover operator is the same as proposed by Falkenauer [19] and is modified as follows:

Step 1. The first parents are selected and the point to be a crossover point in each group of them is determined.
Step 2. Genes belonging to the selected group are copied from the first parent to child.
Step 3. Genes belonging to the second person are copied if it is not specified in the first one.
Step 4. A group is determined randomly for the genes that do not belong to a group.
Step 5. The cluster is deleted if it is empty.
Step 6. Finally members of the chromosomes of child are renamed.

To search properly, the probability of crossover changing dynamically is defined as follows:

$$p_c(k) = p_{ci} + \frac{j}{TG}(p_{ci} - p_{cf})$$

In the above equation, $p_c(k)$ is the probability of crossover used in k-th generation and TG fixed for the total number of generation in algorithm that is determined by repeating the test p_{ci} and p_{cf} as considered the initial and final values for probability.

Example1: Two chromosomes randomly selected as a parent and then from each group section, two clusters are randomly selected:

First parent = [1 3 2 1 4 1 3 4 2 1|1 **2 3** 4]
Second parent = [3 1 2 1 3 2 3 2 2 2|1 **2 3** 4]
To begin with, the identified genes are copied from the first parent to child:
child = [- 3 2 - - 3 - 2-|2 3]

If the genes identified by the second parent are not specified in the first, the copy is for children. Since both parents have clusters 1 and 2, the names of clusters of second parent, 2 and 1 are changed to * 2 and * 1 for the chromosome child:
The child = [- 3 2 2* -2* 3 – 2 2*| 2 3 2* 1*]
Genes that does not belong to a cluster are randomly completed with one of the new clusters and then if there are empty clusters it is removed from the group:
The child = [2* 3 2 2* 1* 1* 3 2 2 2* |2 32* 1*]
Finally the child chromosomes according to the new numbering are arranged and instead of * 1and * 2 replaced 3and 4 their genes are changed to the new values of 3, 4.
The child = [32 1 3 3 4 2 1 1 4 | 1 2 3 4]
For cases when the algorithm suffers local optimal, it helps algorithm to find new areas to search with the changes in each of the chromosomes with low probability. Two different types of mutation were studied:
A) Mutation by gaping clusters: One cluster in this way is selected for mutation. Clusters with more genes have more chance of being selected for mutation. Genes of clusters thus chosen are divided into two categories; one of the categories is in the initial cluster while another is named with n + 1, where n shows the total number of clusters. For example, consider the following chromosomes:
p = [**3** 2 **1 3 1 3 1** | **1** 2 **3**]
In this chromosome, the third cluster is split in two. Data belonging to the third cluster x_1, x_4 and x_6 are with equal probability belong to two groups: x_1 is in the third cluster and x_6 and x_4 are in the new cluster.
Because the total number of clusters is three, the new cluster is named fourth cluster. Chromosome child is changed as follows:

C = [**3** 2 **1 4 1 4** 1 | **1** 2 3 **4**]

B) Mutations by combining clusters: Unlike previous methods here data belonging to two clusters are replaced in one cluster. In this method clusters are also chosen and cluster with a larger size has more chance to be selected. Assume the following chromosome mutation is selected:

P = [**3 2 1 3 1 3 1** | **1 2 3**]

With the mutation data clusters of two and three are replaced in the cluster with smaller number, that means two.

C = [**2 2 1 2 1 2 1** | **1 2**]

The probability mutation is also dynamically considered as:

$$p_m(k) = p_{mi} + \frac{j}{TG}(p_{mf} - p_{mi})$$

Here $p_m(k)$ is the probability of mutation in the k-th generation and TG is a fixed value for the number of total generations and the probable values for the first generation and last generation are p_{mi} and p_{mf}, respectively. The best values for p_{mi} and p_{mf} are considered. These are selected with repeated use and testing.
Replacement and elitism: Crossover and mutation are applied to form a new population to be replaced with current population. The algorithm identifies the elite in each population and transmits to next generation. This is achieved quickly to the optimal solution.
Local search: In this algorithm with a small probability p_h for each individual search is performed on the individual element. As long as the optimal value for the objective function is not reached, the individual chromosomes are modified; the operator used here is likely the crossover operator.
Island model: to improve the efficiency of the algorithm; the model has islands; each island tries to get the best generation. Elite people in each island can move freely between the islands. In fact, migration is performed elitist between the islands, this causes reaching to a solution faster.
The migration process is done in three steps:
1. Selection of people who have most fitness in each island.
2. A destination island is randomly selected for an elite migrating to it.
3. One person is randomly selected to replace with elite in Destination Island [24].

Example2: We want to divide workers between three parts of a factory.

For this division there is a restriction, such as the coordination of education and jobs to the area the skill of the worker and some other items have been tested. Let studied chromosomes to be as follows:
[1 2 1 1 1 2 2 1 3 3 |1 2 3]
Three parts of factories are grouped in our question. Local search for each worker calculates fitness chromosomes when the workers belong to other groups. The object is to place it in the best group which has the best fitness function value. The amount of fitness of each worker for three groups is calculated as shown in Table 1:

Worker	Group1	Group2	Group3
#1	65	80	9
#2	16	32	41
#3	62	34	10
#4	82	12	16
#5	74	10	19
#6	13	43	8
#7	76	20	4
#8	69	46	9
#9	53	11	25
#10	78	8	84

Table 1: The fitness of workers in different parts of factory

According to the Table1, for the first worker, the fitness function for the first group is 65, second group is 80 and third group is 9. So the best group that first worker could belong to it is the second group. Regarding second worker, the fitness function when placed in the first group is 16, the second group is 32 and the third group is 41, so the best group that second worker could belong to it is the third group. While the second worker in the chromosomes presented above, is in the second group; this must be corrected. This changing continues so on. At the end the chromosome is changed in the following format:

[1 2 1 1 1 2 2 1 3 3 |1 2 3] →[1 3 1 1 1 2 1 1 1 3 |1 2 3]

Search for all chromosomes must be done with a small probability. Each chromosome is evaluated when its genes are moved.

5. A new way to improve Grouping Genetic Algorithm

In many meta-heuristic algorithms such as Genetic Algorithms that initial solution is randomly selected, there is the possibility of production of local optimal, which prevent them from reaching the optimal solution or slows down the process of reaching an answer.

To fix this problem and speed up to reaching solution, we try to combine Grouping Genetic Algorithm and DBSCAN, to present a new algorithm for clustering.

5.1 The combination of Grouping Genetic Algorithm and DBSCAN

We use a combination of clustering method. This means that in the problem of clustering, DBSCAN method is used for clustering the initial population. Then the algorithm uses the results to obtain the final and optimal clustering.

Encoding: chromosomes encoded similar to GGA algorithm. From now on the order of GGA is Grouping Genetic Algorithm and IMGGA, Improved Grouping Genetic Algorithms which is proposed for clustering problems. The results of implementation with regard to the fixed length for the chromosomes have been investigated.

Use K-mean algorithm for finding cluster centers: Based on the chromosomes, each center of clusters are computed. Consider the following chromosome that is randomly generated:

[1 2 2 1 1 2 2 2 3 3 1 2 1 3 1|1 2 3]

As a result, the clustering of chromosomes is as follows:

$C_1 = [x_1, x_4, x_5, x_{11}, x_{13}, x_{15}]$, $C_2 = [x_2, x_3, x_6, x_7, x_8, x_{12}]$, $C3 = [x_9, x_{10}, x_{14}]$

Suppose data are two-dimensional as follows:

x_1 =(1.4,6.7) , x_2=(10.4, 3) ,x_3=(11,1.5) , x_4= (8 , 2) , x_5=(1,1)

x_6=(9 , 9) , x_7=(1.3 ,4.3) , x_8=(3,3) , x_9=(0,1.2) x_{10}= (5.2 ,7.1)

x_{11}=(6 ,2) , x_{12}=(6.5 , 1) , x_{13}=(2 ,2) , x_{14}=(5,1) , x_{15}=(18, 12)

Average data are calculated for each cluster and are considered as cluster centers. The result of the cluster centers is as follows:

Average of cluster 1: (6.06, 4.28)

Average of cluster 2: (6.86, 3.63)

Average of cluster 3: (3.4, 3.1)

The new center replaces the old center.

Use K-mean algorithm for modifying chromosome results: For each chromosome distance data from each cluster centers are calculated. Data is placed in the cluster that is closest to the center.

Selection operation: For selecting the parent we can choose the Roulette wheel with a selection of elite persons to choose parents to retain the best people in each generation. This method improves the selection process and saves the

best ones. By choosing elite, the best solution in each generation increases with time evenly. If we do not use elitist selection, random errors may make people disappear.

Crossover: The crossover operator is the same as proposed by Falkenauer.

Mutation: Single-point mutation is used.

Replacing: The population will replace current population.

The stopping criterion: Algorithm stops after a defined number of generations.

6. Experiments

Experiments on Iris data sets, wine and the balance by GGA and IMGGA using RAND index have been studied and the results are shown in the following Table 2:

algorithm	Data set	Number of repetitions
GGA	Iris	132
	Wine	147
	Balance	118
IMGGA	Iris	97
	Wine	88
	Balance	85

Table 2: Experiments result on Iris data sets, wine and the balance by GGA and IMGGA

Evaluation charts of GGA and IMGGA for grouping data of above Table can be observed in the following Figures:

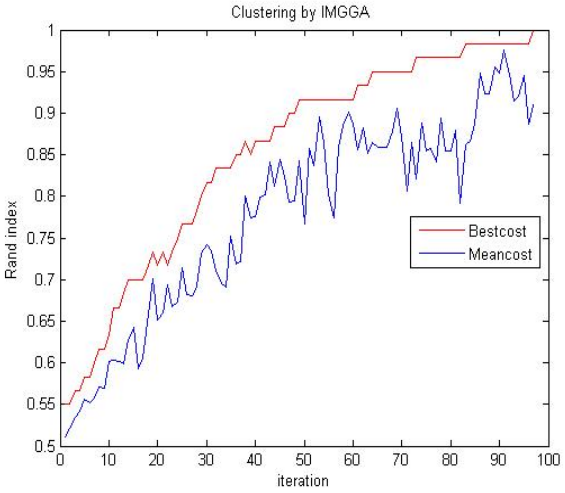

Figure1: Evaluation charts of GGA for iris data

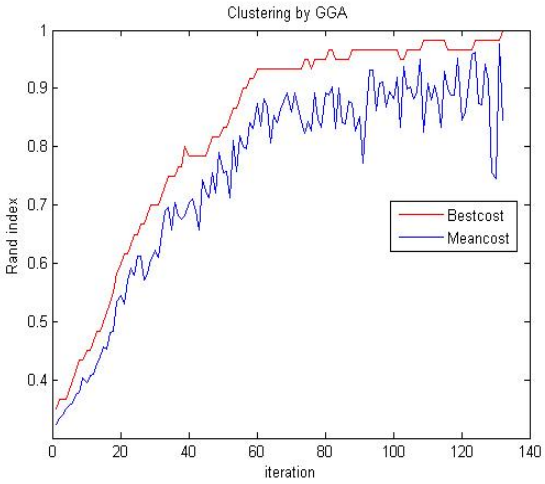

Figure2: Evaluation charts of IMGGA for iris data

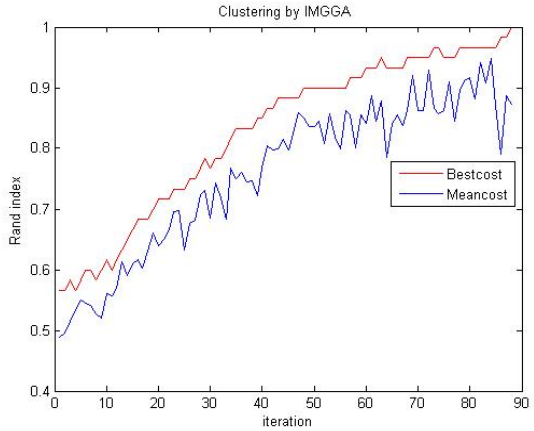

Figure3: Evaluation charts of GGA for wine data

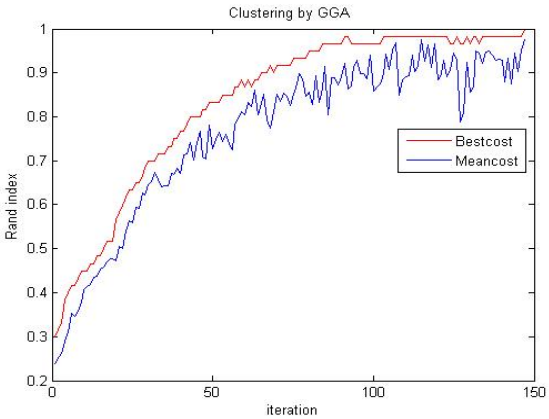

Figure4: Evaluation charts of IMGGA for wine data

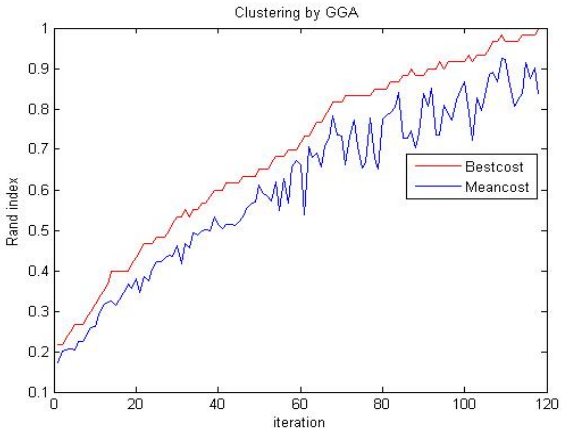

Figure5: Evaluation charts of GGA for balance data

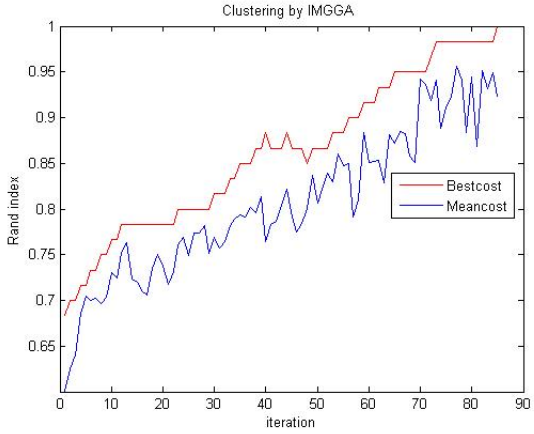

Figure6: Evaluation charts of IMGGA for balance data

The results show that although both cases get the index value as the final results, but as it can be observed IMGGA achieves the optimal solution faster with fewer numbers of iterations generally. The results show that in 70% of cases, the proposed algorithm is better than the GGA. In fact, the improved algorithm can achieve the optimal solution more quickly.

7. Conclusion

This paper introduced several efficient methods in clustering problems. These techniques include:

A) Grouping Genetic Algorithm

B) Grouping Genetic Algorithm for clustering problem

Grouping Genetic Algorithm is derived from Classical Genetic Algorithm for clustering problems and improved the previous model.

In improved clustering algorithm, a perfect description of operators and other implementation details, such as local search was presented. The idea of island was introduced to improve the implementation of the algorithm. Finally, a proposed algorithm for clustering was presented and its performance was tested and the results showed good performance in order to reach the optimal solution.

REFERENCES

[1] Murthy, C. A., Chowdhury, N., 1996. In search of optimal clusters using Genetic Algorithms, Pattern Recognition Lett.17: 825-832.

[2] Falkenauer, E., 1999. Genetic Algorithms and Grouping Problems, Wiley, New York, NY.

[3] Bandyopdhyay, S., Maulik, U., 2002. An evolutionary technique based on K-means algorithm for optimal clustering in RN, Inf.Sci.146: 221-237.

[4] Bandyopdhyay, S., Saha, S., 2007. GAPS: a clustering method using a new point symmetry-based distance measure, Pattern Recognition40: 3430-3451.

[5] Krishna, K., Murty, M.N., 1999. Genetic K-means algorithm, IEEETrans. Syst. Man Cybern. PartBCybern.29: 433-439.

[6] Maulik, U., Bandyopadhyay, S., 2000. Genetic algorithm based clustering technique, Pattern Recognition33: 1455-1465.

[7] Scheunders, P., 1997. A genetic c-means clustering algorithm applied to color image quantization, Pattern Recognition 30 (6): 859-866.

[8] Hall, L. O., Ozyurt, I. B., Bezdek, J. C., 1999. Clustering with a genetically optimized approach, IEEETrans. Evol. Comput.3 (2): 103-112.

[9] Tseng, L. Y., Yang, S.B., 2001. A genetic approach to the automatic clustering problem, Pattern Recognition34 (2): 415-424.

[10] Laszlo, M., Mukherjee, S., 2006. A Genetic Algorithm using hyper-quad trees for low- dimensional K-means clustering, IEEE Trans. Pattern Anal. Mach.Intell.28 (4): 533-543.

[11] Laszlo, M., Mukherjee, S., 2007. A Genetic Algorithm that exchanges neighboring centers for K-means clustering, Pattern Recognition Lett. 28: 2359-2366.

[12] Fleurya, G., Hero, A., Zareparsi, S., Swaroop, A., 2004. Gene discovery using Pareto depth sampling distributions, J. Franklin Inst. 341 (1–2): 55-75, (special number on Genomics, Signal Processing and Statistics).
[13] Morita, M., Sabourin, R., Bortolozzi, F., Suen, C.Y., 2003. Unsupervised feature selection using multi-objective Genetic Algorithms for handwritten word recognition, in: Proceedings of the 7th International Conference on Document Analysis Recognition, 666-671.
[14] P. Gancarski, P., Blanscheé, A., Wania, A., 2008. Comparison between two coevolutionary feature weighting algorithms in clustering, Pattern Recognition 41: 983-994.
[15] Chang, D., Zhao, Y., Zheng, C., & Zhang, X., 2012. A Genetic Clustering Algorithm using a message-based similarity measure. Expert Systems with Applications, 39(2): 2194-2202.
[16] Asuncion, A., & Newman, D. J., 2007. UCI Machine Learning Repository, http://www.ics.uci.edu/mlearn/MLRepository.html. Irvine, CA: University of California, School of Information and Computer Science.
[17] Melanie, M., 1999. An Introduction to Genetic Algorithms, *Massachusetts Institute of Technology*, Cambridge, Massachusetts • London, England.
[18] Hruschka, E., Hruschka, E., Ebecken, N., 2005. Missing Values Imputation for a Clustering Genetic Algorithm, Proceedings Of The First International Conference On Advances In Natural Computation, 3: 245-254.
[19] Falkenauer, E., 1992. The Grouping Genetic Algorithm widening the scope of the GAs. In Proceedings of the Belgian Journal of operations research, statistics and computer science 33: 79-102.
[20]. Falkenauer, E., , 1997. Genetic Algorithms and Grouping Problems, John Wiley & Sons, New York.
[21] Brown, E. C., Sumichrast, R. T., 2001. CF-GGA: A Grouping Genetic Algorithm for the Cell Formation Problem, *International Journal of Production Research,* 39 (16): 3651-3669.
[22] Sivara, R., Ravichandran, T., 2011. An Efficient Grouping Genetic Algorithm, *International Journal of Computer Applications*.
[23] Bäck, T. and Hoffmeister, F., 1991. Extended Selection Mechanisms in Genetic Algorithms, Proceedings of the Fourth International Conference on Genetic Algorithms, 92-99.
[24] Agustin-Blas, L.E., Salcedo-Sanz, S., Jimenez-Fernandez, S., 2012. A New Grouping Genetic Algorithm For Clustering Problems, *Expert Systems with Applications*, 39: 9695-9703.
[25] James, T. L., Brown, E. C., & Keeling, K. B., 2007. A hybrid Grouping Genetic Algorithm for the cell formation problem, Computers & Operations Research, 34: 2059-2079.