British Journal of Mathematics & Computer Science
6(4): 336-350, 2015, Article no.BJMCS.2015.082
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Data Clustering Using Hybridization Strategies of Continuous Ant Colony Optimization, Particle Swarm Optimization and Genetic Algorithm

Moein Fazeli Hassan Abadi[1*] and Hassan Rezaei[1]

[1]Department of Computer Science, Faculty of Mathematics, University of Sistan and Baluchestan, Iran.

Article Information

DOI: 10.9734/BJMCS/2015/15341
Editor(s):
(1) Dariusz Jacek Jakóbczak, Computer Science and Management Department, Technical University of Koszalin, Poland.
Reviewers:
(1) Anonymous, China.
(2) Anonymous, India.
(3) Francisco Beraldo Herrera Fernandez, Departamento de Automática y Sistemas Computacionales, Universidad Central "Martha Abreu" de Las Villas, Cuba.
(4) Anonymous, India.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=734&id=6&aid=7677

Original Research Article

Received: 21 November 2014
Accepted: 29 December 2014
Published: 09 January 2015

Abstract

Nowadays, clustering plays a critical role in most research areas such as engineering, medicine, biology, data mining, etc. Evolutionary algorithms, including continuous ant colony optimization, particle swarm optimization, and genetic algorithms, have been employed for data clustering. To improve searching skills, this paper examines four strategies, combining of continuous ant colony optimization and particle swarm optimization, and proposes a strategy which is a combination of these two algorithms with genetic algorithm. Available methods and the proposed method were implemented over several sets of benchmark data to assess the validity. Results were compared with the results of continuous ant colony optimization and particle swarm optimization. The high capacity and resistance of combined methods are obvious according to results.

Keywords: Data clustering, continuous ant colony optimization, particle swarm optimization, genetic algorithm.

1 Introduction

Clustering is a process of grouping a set of objects into clusters. The internal members of each cluster look very similar to each other and are not similar to the members of other clusters.

**Corresponding author: moein_fazeli@pgs.usb.ac.ir;*

Clustering is employed in different sciences such as engineering, medicine, social sciences, and marketing. It is of data mining techniques, which extracts models from data without a pre-defined purpose. Certain popular clustering methods can be adopted such as partitioning methods, hierarchical methods, grid-based methods, model-based methods, density-based methods [1] and fuzzy Clustering [2].

To analyze clusters, objects are displayed in a n-dimensional space. Vectors show the properties of objects. The main purpose is to classify n objects into k clusters with objects looking very similar to each other. This paper uses Euclidean distance to minimize the deviation of internal points of clusters from a central point.

Various techniques and algorithms are employed to cluster data. Most problems relate to convex and nonlinear clustering with local minimum answers. As a result, an optimal answer may not be found. In algorithms in which the primary answer is randomly selected, a local optimal answer may be found and prevent from reaching an optimal answer.

Swarm intelligence, including continuous ant colony optimization and particle swarm optimization, is used in clustering. Ant-based clustering was first introduced by Deneubourg et al. Particle swarm optimization and genetic algorithms have been used in many research studies [3-9]. Studies disclosed that by advancing in this area, these methods can prevent from falling into local optimum and have better performance than some other traditional clustering algorithm. Combining algorithms can improve the main algorithm and give a higher qualified answer. In evolutionary calculations, combination is necessary to improve the optimization algorithm. The performance of an alone algorithm is lower than combined algorithms. In the present algorithm, several new and effective methods have been developed by combining metaheuristic algorithms of continuous ant colony optimization [10], particle swarm optimization [6] and genetic algorithm [7].

The remainder of our paper is organized as follows: section 2 reminds basic concepts, including continuous ant colony optimization (ACOR), particle swarm optimization (PSO), data clustering and Genetic Algorithm (GA); section 3 explains combined systems ACOR-PSO, section 4 explains combined system ACOR-PSO-GA; The computational complexity of the hybrid model described in section 5; section 6 presents the empirical results of UCI dataset, and conclusions are described in Sections 7.

2 Preliminaries

2.1 Continuous Ant Colony Optimization

This algorithm was introduced by Sosha and Dorigo [10] to solve optimization problems for continuous functions. According to this method, any line of Pheromone table presents a solution from a collection of decision-making variables. Each solution includes a value from the target function. In ACOR we keep track of a number of solutions in a pheromone table. For each solution s_l to an n-dimensional problem, ACOR stores in k (k is the number of solutions in pheromone table) the values of its n variables and the value of the objective function $f(s_l)$. The i^{th} variable of l^{th} solution is hereby denoted by s_l^i. The structure of the pheromone table is presented in Fig. 1. In this method, primary solutions are randomly found. The new ants in the next generation are produced by a Roulette Wheel probability achieved based on the target function of each solution of the pheromone table. The algorithm is detailed as follows:

1) The target function f (s_i) is calculated for any solution s_i in the pheromone table. Available solutions in the table are arranged based on the target function. Thus, for a minimization problem, we have**:**

$$f(s_1) \leq \cdots \leq f(s_i) \leq \cdots \leq f(s_i) \leq f(s_k),$$

2) Weight (w) is calculated for i^{th} solution in the pheromone table as below:

$$w_i = \frac{1}{qk\sqrt{2\pi}} e^{-\frac{(i-1)^2}{2q^2K^2}} \tag{1}$$

In the above equation q shows the learning rate, which is a value between 0 and 1.

3) As to w_i for each available solution in the pheromone table, Roulette Wheel probability p_i is estimated as bellow:

$$p_i = \frac{w_i}{\sum_{j=1}^{k} w_j} \tag{2}$$

4) Stage 4 is iterated M times to produce M new ants (M ≤ K): using normal distribution(μ_i^d, σ_i^d), a new ant is achieved for each variable. μ_i^d isa value selected from d^{th} variable and from i^{th} solution in the pheromone table with the probability p_i. σ_i^d is defined as follows:

$$\sigma_i^d = \tau \sum_{j=1}^{k} \frac{\left|x_j^d - x_i^d\right|}{K-1} \tag{3}$$

In the above equation, x_i^d is the d^{th} variable from the i^{th} solution. K is the size of the pheromone table and τ shows the rate of evaporation which is between 0 and 1.

5) M new ants are evaluated and less qualified solutions are replaced in the pheromone table by superior solutions and by M new ants.

2.2 Particle Swarm Optimization

It was developed by Kennedy and Eberhart, in 1995 and has been successfully employed in many scientific and applied areas. This is a population-based algorithm in which anyone is considered as a particle and any population consists of a number of these particles. In PSO, problem-solving space is regarded as a searching environment and anyplace is a problem-dependent solution. In this population, particles tries to find the best situation (best solution) in the searching place (solution space). All particles move according to their speed. The movement of particles in any iteration is calculated by the following formula:

$$v_i^d(t) = w \times v_i^d(t) + c_1 \times rand_1 \times (pbest_i^d - x_i^d(t-1)) + c_2 \times rand_2 \times (gbest^d - x_i^d(t-1)) \tag{4}$$

$$x_i^d(\mathrm{t}) = x_i^d(\mathrm{t}-1) + v_\mathrm{i}^\mathrm{d}(\mathrm{t}) \tag{5}$$

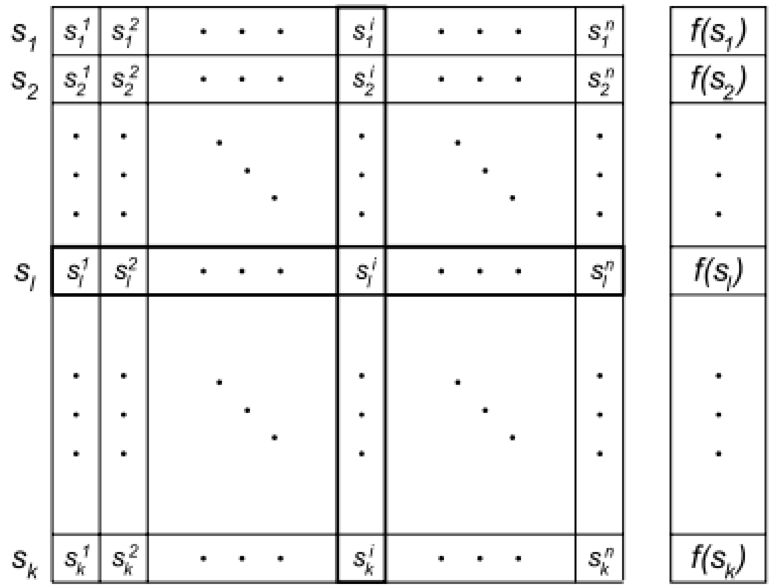

Fig. 1. The structure of the pheromone table

In relations (4) and (5), x_i^d is the current position of d^{th} dimension of i^{th} particle and $v_i{}^d$ is the current speed of this dimension of this particle, and $pbest_i^d$ is the previous optimal position of the d^{th} dimension encountered by i^{th} particle, $gbest^d$ is the best position of d^{th} dimension which has been ever found by the population, w is the weight of inertia that gives a proportion of the previous speed,c_1 and c_2 are the acceleration coefficients and define the best effect of the position of each particle and the best overall position, $rand_1$ and $rand_2$ are two random numbers between 0 and 1. Fig. 2 presents the procedure of particle swarm optimization [3,6].

2.3 Genetic Algorithm

The genetic algorithm is an optimization technique and a random search method which has been produced by the concepts of natural selection theory and evolutionary processes. According to Darwin's evolutionary theory, in nature, some people of a population who are more qualified for living are survived and others perish in a competition for living [1]. Two basic operators of genetic algorithm are crossover and mutation. The applied cutting operator is a two-point crossover and the act of mutation occurs accidentally over a gene, Fig. 3 shows the details of crossover and mutation. Fig. 4 displays the procedure of genetic algorithm.

2.4 Solution Structure and Objective Function

In continuous ant colony optimization, solutions, called pheromones, are recognized in particle swarm optimization as particles and in genetic algorithm as chromosome, which have similar structures. One vector solution is from real numbers in dimensions of k × d. k is the number of clusters and d is the data dimensions, which should be clustered. Fig. 5 presents a sample of this solution. All variables are standardized before clustering. In this study, each variable is mapped in [0,1] by linear normalization. Relation (6) is used in this regard.

Initialize a population of particles with random positions and velocities in the search space.
While(termination conditions are not met)
{
For each particle i do
Update the position of particle i accordingt equation (4).
Update the velocity of particle i according to equation (5).
Map the position of particle i in the solution space and evaluate its fitness value according to the fitness Function.
Update p_i and g_i if necessary.
End for
}

Fig. 2. Procedure of particle swarm optimization

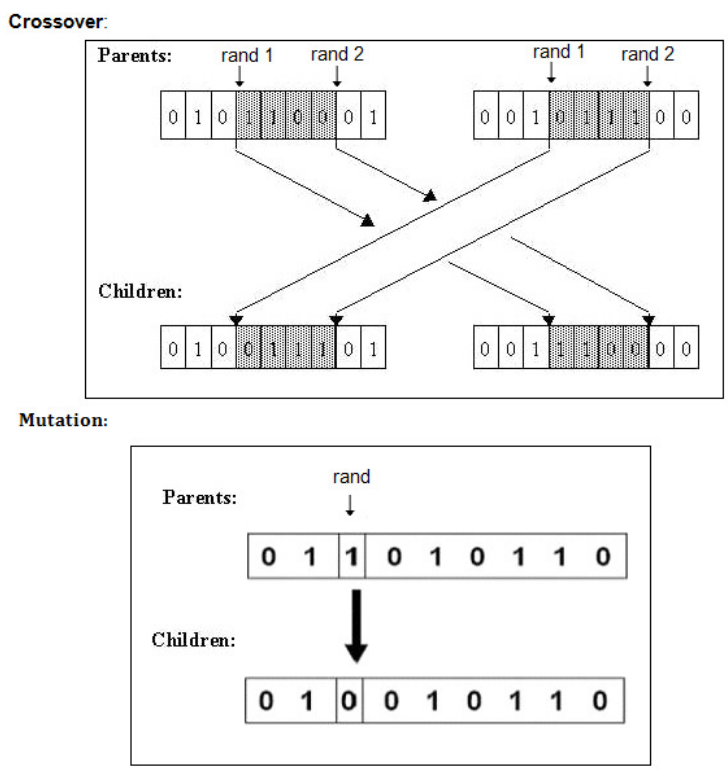

Fig. 3. Crossover and mutation in genetic algorithm

```
Genetic Algorithm
begin
    Choose initial population
    repeat
        Evaluate the individual fit nesses
        of a certain proportion of the population
        Select pairs of best-ranking individuals to reproduce
        Apply crossover operator
        Apply mutation operator
    until terminating condition
end
```

Fig. 4. Procedure of genetic algorithm

X_{11}	X_{12}	…	X_{1d}	…	X_{k1}	X_{k2}	…	X_{kd}

Fig. 5. Structure of a solution

$$x^{\text{new}} = \frac{x - \min(x)}{\max(\mathrm{x}) - \min(\mathrm{x})} \tag{6}$$

In relation 6, x^{new} is the normalized value for *x* and *min (x)* and *max (x)* are the smallest and largest data respectively. The target function which is used to evaluate the fitness of clustering is defined as follows:

$$F = \frac{\sum_{i=1}^{N_s} \min_{k=1}^{N_c} \| x_i - c_k \|^2}{\sum_{k,j=1,k\neq j}^{N_c} d(C_k, C_j)} \tag{7}$$

N_c= number of center

N_s = sample size

$\|x_i - c_k\|^2$ = distance between sample i to center k

$d(C_k, C_j)$ = distance between center k and center j

2.5 Data Clustering

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. We can show this with a simple graphical example in Fig. 6:

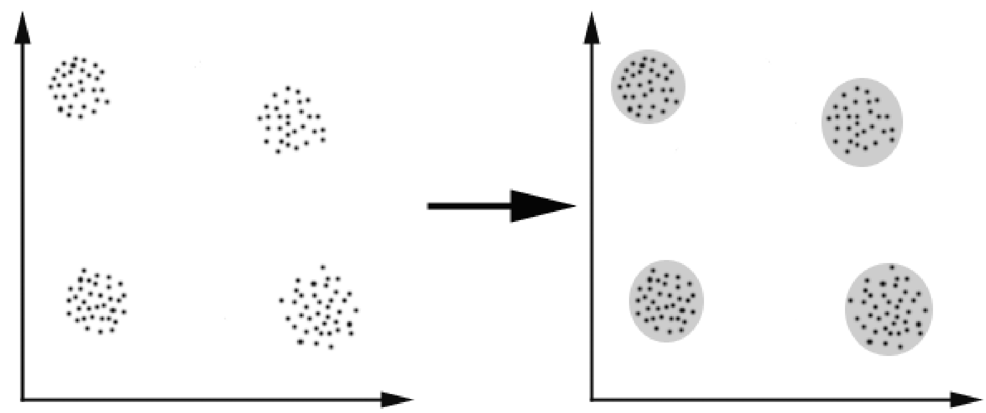

Fig. 6. Example of clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering.

Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

2.5.1 The goals of clustering

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.

For instance, we could be interested in finding representatives of homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

2.5.2 Possible applications

Clustering algorithms can be applied in many fields, for instance:

- Marketing: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- Biology: classification of plants and animals given their features;
- Libraries: book ordering;
- Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;

- City-planning: identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: clustering observed earthquake epicenters to identify dangerous zones;
- WWW: document classification; clustering weblog data to discover groups of similar access patterns.

3. Hybridization Strategies of Continuous Ant Colony Optimization and Particle Swarm Optimization

This section explains four combined models created by combination of the Continuous Ant Colony Optimization and particle swarm optimization.

3.1 Sequential Approach

In this method, ACOR and PSO share a set of answers with each other, which is called "pheromone-particle table". This set of answers includes pheromone in ACOR and current solution in PSO. Based on pheromone-particles table, PSO produces newer particles and replaces lower particles and solutions in pheromone-particles table with solutions superior to new particles. According to pheromone- particles table updated by PSO, ACOR produces new ants and replaces lower particles and solutions in pheromone-particles with solutions superior to new ants. The features of the sequential method include: (1) superior solutions produced by PSO and ACOR can be kept in pheromone table; (2) ACOR produces new ants and replaces low solutions in pheromone table. This brings variety in the pheromone table and prevents from local optimization. Fig. 7 shows the main stages of the sequential approach [10]. In this method, if the first run ACOR and then run PSO, the final solutions not difference.

3.2 Parallel Approach

In this method, According to pheromone-particles table, PSO produces new particles and ACOR produces new ants. Underneath solutions in pheromone-particles table are replaced by solutions superior of K new particles and M new ants. Fig. 8 shows the main stages of the Parallel approach [10].

3.3 Sequential Approach with the Enlarged Pheromone-Particle Table

In this method, PSO produces k new particles based on pheromone-particles table. These particles are combined with pheromone-particles table and form an extended table with 2k in size. In PSO, pbest and gbest are achieved from updated pheromone-particles table. In ACOR, ants are extended based on pheromone tables with varied production. This method prevents from local optimum. Fig. 9 shows the main stages of this approach [10].

3.4 Global Best Exchange

In global best exchange method, PSO produces new particles based on its particle table. ACOR produces new ants based on its pheromone table. Two models exchange their best solutions. This approach keeps the main features of PSO and ACOR. Fig. 10 shows the main stages of this approach [10].

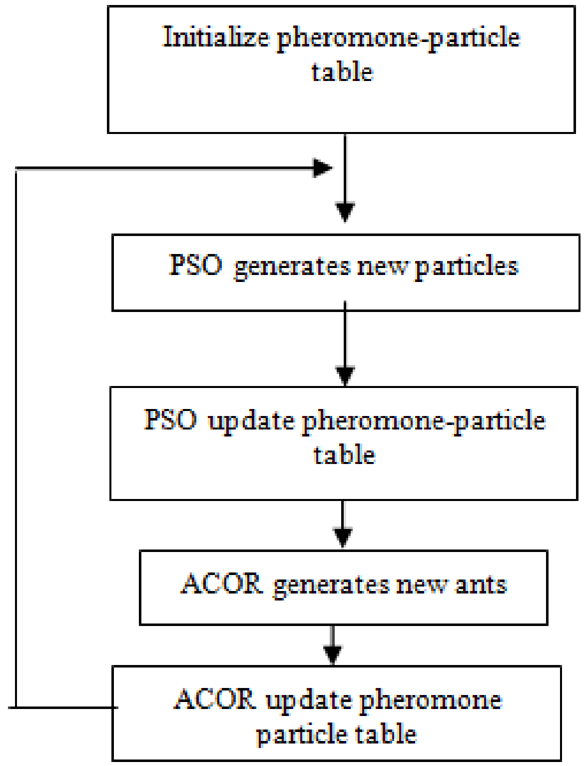

Fig. 7. Sequential approach

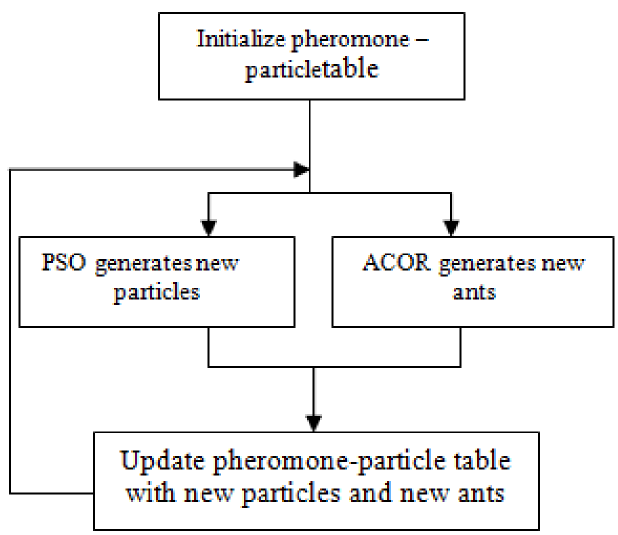

Fig. 8. Parallel approach

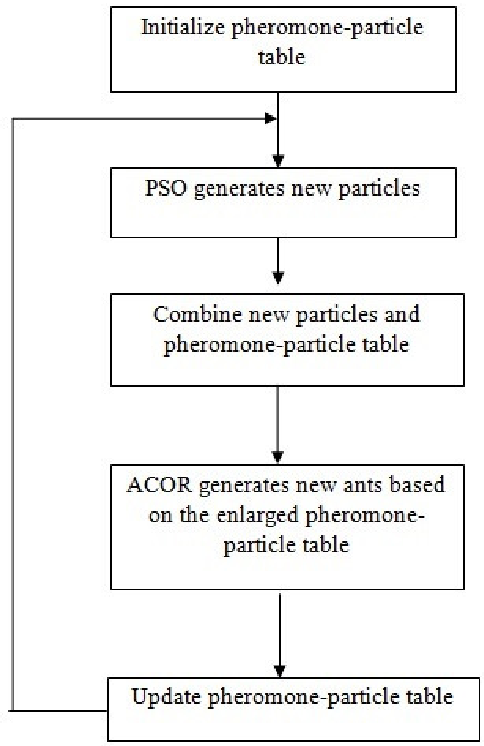

Fig. 9. Sequential approach with the enlarged pheromone-particle table

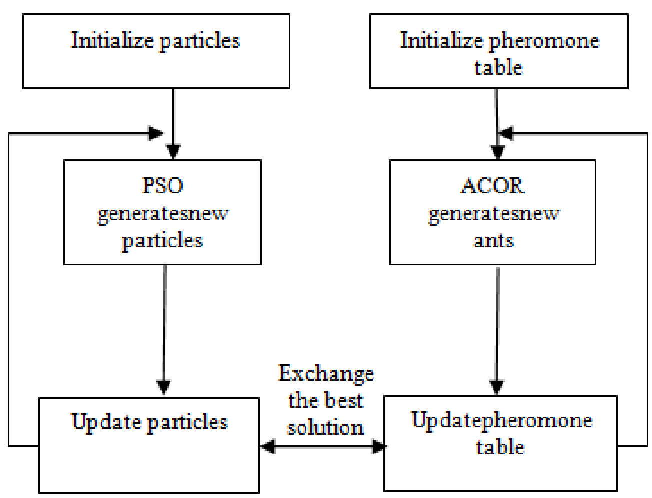

Fig. 10. Global best exchange approach

4. Proposed Model

In such combined strategy, ACOR, PSO and genetic algorithm share a set of answers with each other which is called pheromone-particles-chromosome table. This set of answers includes pheromone in ACOR, current solutions in PSO and chromosomes in genetic algorithm. Based on pheromone-particles-chromosome, PSO produces new particles and replaces lower particles and solutions in pheromone-particles-chromosome table with solutions superior to new particles. According to pheromone- particles-chromosome table updated by PSO, ACOR produces new ants and replaces lower particles and solutions in pheromone-particles-chromosome table with solutions superior to new ants. Then, based on this table, genetic algorithm produces new chromosome. Lower solutions in pheromone-particles-chromosome table are then replaced by superior solutions of new chromosomes. The features of this method include: (1) superior solutions produced by PSO, ACOR and GA can be kept in solution table; (2) ACOR produces new ants and replaces low solutions in solution table; (3) GA produces new chromosomes and replaces low solutions in solution table. This brings variety in the solution table and prevents from local optimization. Fig. 11 shows the main stages of this approach.

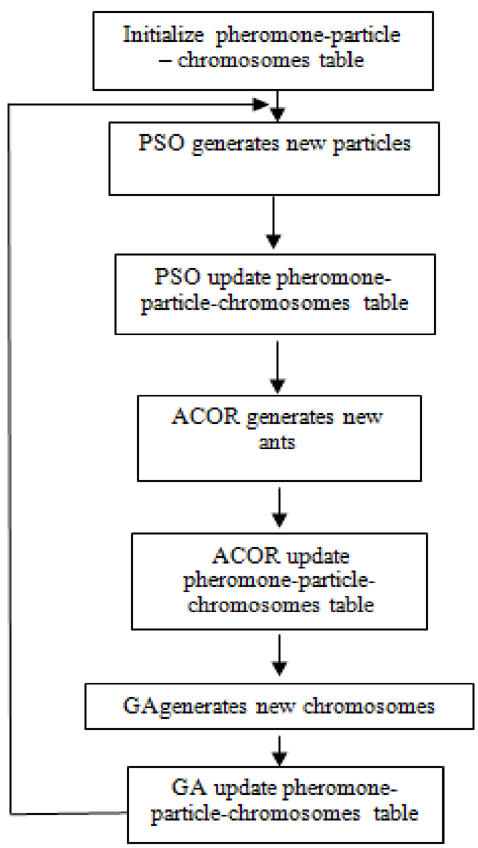

Fig. 11. Hybridization strategies of continuous ant colony optimization, particle swarm optimization and genetic algorithm

5. Computational Complexity Analysis

The time complexity of the hybridization strategies of the ACOR-PSO are Discussed in [3], and the time complexity of the proposed model as follows:

$$N_{iterations} \times \left(N_{NewAnts} \times \mathrm{T_{Ant}} + N_{New\mathrm{Particle}} \times \mathrm{T_{Particle}} + T_{New\mathrm{Chromosome}} \times N_{New\mathrm{Chromosome}} + \mathrm{T}_{Update}\right) \tag{8}$$

Where:

$\mathrm{N_{iterations}}$: Number of iterations
$\mathrm{N_{NewAnts}}$: Number of new ants
$\mathrm{N_{NewParticle}}$: Number of new particles
$\mathrm{N_{NewChromosome}}$: Number of new chromosome
$\mathrm{T_{Ant}}$: Runtime for generating a new ant
$\mathrm{T_{Particle}}$: Runtime for generating a new particle
$\mathrm{T_{NewChromosome}}$: Runtime for generating a new chromosome
$\mathrm{T_{Update}}$: Runtime for update pheromone-particles-chromosome table

6. Experiments

Empirical results in [3] show that Sequential Approach with the enlarged pheromone-particle table are better than three other combined methods of PSO and ACOR, standalone ACOR and standalone PSO.

In this study the proposed model, DCPG [11] algorithm and sequential approach with the enlarged pheromone-particle table with eight dataset from UCI dataset tested for clustering to verify the accuracy of the proposed model.

The proposed implementation was executed on the Matlab9 platform, with an Intel Core 2 Duo CPU running at 3.0GHz and 4GB RAM. Table 2 presents parameters required to implement algorithms.

In Table 1 show dataset of UCI. Relation (7) displays the efficiency of clustering. The average answers of current algorithms over eight dataset from UCI dataset show in Tables 3, 4 and 5. To summarize the sequence with enlarged pheromone-particle table model was introduced as hybrid I and ACOR-PSO-GA model (proposed model) as hybrid II.

Table 1. Datasets from the UCI repository

No	Names	#Instances	Numeric features
1	Iris	150	3
2	Wine	178	13
3	Contraceptive method choice (CMC)	1473	9
4	Liver Disorders	354	6
	German (credit card)	1000	24
6	ISOLET	7797	617
7	Turkiye Student Evaluation	5820	33
8	Letter Recognition	20000	16

Table 2. System parameter setting

Parameter	Hybrid I	DCPG algorithm	Hybrid II
Learning rate C1	1.5	1.49	1.5
Learning rate C2	1.5	1.49	1.5
Learning rate q	1.5		0.95
Evaporation rate	1.5		0.05
Number of particles	20	20	20
Size of Pheromone-particle table	15		15
Number of ants	5		5
Number of iterations	100	100	100
Number ofchromosomes		20	20
Size of Pheromone-particle- chromosomes table			30
Mutation rate		1	0.4
Crossover rate		0.05	0.05

Table 3. The averages answer in four runs of hybrid I over datasets

Dataset name	The average cost of the best solution in four runs	Average time in four runs (min)
Iris	0.53655	2.4
wine	0.512925	2.51
CMC	7.6416	4.02
Liver Disorders	1.7849	2.55
German (credit card)	23.3683	3.9517
ISOLET	205.147	11.533
Turkiye Student Evaluation	232.158	26.512
Letter Recognition	322.04	27.692

Table 4. The averages answer in four runs of DCPG algorithm over datasets

Dataset name	The average cost of the best solution in four runs	Average time in four runs (min)
Iris	0.6587	2.08
wine	0.8414	2.43
CMC	7.98245	3.82
Liver disorders	2.25	2.45
German (credit card)	25.36	3.657
ISOLET	207.124	10.543
Turkiye Student Evaluation	231.158	24.512
Letter Recognition	328.04	26.2

Empirical results show that combined methods are effective and hybrid II is more efficient than Hybrid I and DCPG algorithm. In this model, superior solutions in iterations are kept. Since three different algorithms are used in this method, there is a wide variety of population, as it is improbable for a global optimization to occur.

One possible reason for superior performance of the Hybrid II is that the use of GA that improves the diversity for generating new solutions. The diversity of new populations can improve the solution quality [12]. Some possible strategies may improve the diversity of populations including Sub-populations, communication and migration between sub-populations [13], random velocity in

PSO and mutation in GA. The diversity of pheromone-particle-chromosome table is another possible way investigated in this study to improve the diversity of populations.

Table 5. The averages answer in four runs of Hybrid II over datasets

Dataset name	The average cost of the best solution in four runs	Average time in four runs (min)
Iris	0.2057	2.42
wine	0.5119	2.52
CMC	4.3351	4.10
Liver disorders	0.9739	3
German (credit card)	19.6252	4.0827
ISOLET	196.6745	12.014
Turkiye student evaluation	229.8634	27.015
Letter recognition	284.1016	28.676

7. Conclusion

To improve data clustering, four combined strategies were used. These strategies were created by combining continuous ant colony optimization and particle swarm optimization. Another combined method was also proposed, which is a combination of these two algorithms with genetic algorithm. Simulation results revealed that the proposed method is preferred to other combined methods, PSO and ACOR.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Han J, Kamber M. Data mining concepts and techniques (2nd ed.), Morgan Kaufmann Publishers, San Francisco, USA; 2006.

[2] Chen KY, Kwong CK, Hu BQ. Market segmentation and ideal point identification for new product design using fuzzy data compression and fuzzy clustering methods, Applied Soft Computing. 2012;12(4):1371-1378.

[3] Chuang LH, Hsiao CJ, Yang CH. Chaotic particle swarm optimization for data clustering. Expert Systems with Applications. 2011;28(12):14555-14563.

[4] Nanda SJ, Panda G. Automatic clustering algorithm based on multi-objective Immunized PSO to classify actions of 3D human models. Engineering Applications of Artificial Intelligence. 2013;26(5-6):1429-1441.

[5] Yu H, Xiaohui W. PSO-based energy-balanced double cluster-heads clustering routing for wireless sensor networks. Procedia Engineering. 2011;15:2073-3077.

[6] Esmin AA, Matwin S. Data clustering using hybrid particle swarm optimization. Intelligent Data Engineering and Automated Learning. 2012;17435:159–166.

[7] Liu Y, Wu X, Shen Y. Automatic clustering using genetic algorithms. Applied Mathematics and Computation. 2011;218(4):1267-1297.

[8] Alam S, Dobbie G, Koh YS, Riddle P, Rehman SU. Research on particle swarm optimization based clustering: A systematic review of literature and techniques. Swarm and Evolutionary Computation. 2014;17:1-13.

[9] Beghi A, Cecchinato L, Cosi G, Rampazzo M. A PSO-based algorithm for optimal multiple chiller systems operation. Applied Thermal Engineering. 2012;32:31-40.

[10] Huang CL, Huang WC, Chang HY, Yeh YC, Tsai CY. Hybridization strategies for continuous ant colony optimization and particle swarm optimization applied to data clustering. Applied Soft Computing. 2013;13(9):3864-3872.

[11] Kuo RJ, Syu YJ, Zhen-Yao Chen, Tien FC. Integration of particle swarm optimization and genetic algorithm for dynamic clustering. Information Sciences. 2012;195:124-140.

[12] Lozano M, Garc A-Mart Nez C. Hybrid metaheuristics with evolutionary algorithms specializing in intensification and diversification: Overview and progress report. Computers & Operations Research. 2010;37(3):481-497.

[13] Twomey C, Stutzle T, Dorigo M, Manfrin M, Birattari M. An analysis of communication policies for homogeneous multi-colony ACO algorithms. Information Sciences. 2010;180(12):2390-2404.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=734&id=6&aid=7677