A Hybrid Discrete Artificial Bee Colony - GRASP Algorithm for Clustering

Yannis Marinakis[1], Magdalene Marinaki[2], Nikolaos Matsatsinis[1]

[1]Decision Support Systems Laboratory, Department of Production Engineering and Management, Technical University of Crete, 73100 Chania, Greece (marinakis@ergasya.tuc.gr, nikos@ergasya.tuc.gr)
[2]Industrial Systems Control Laboratory, Department of Production Engineering and Management, Technical University of Crete, 73100 Chania, Greece (magda@dssl.tuc.gr)

ABSTRACT

This paper presents a new hybrid algorithm, which is based on the concepts of the Artificial Bee Colony (ABC) and Greedy Randomized Adaptive Search Procedure (GRASP), for optimally clustering N objects into K clusters. The proposed algorithm is a two phase algorithm which combines an Artificial Bee Colony Optimization algorithm for the solution of the feature selection problem and a GRASP algorithm for the solution of the clustering problem. As the feature selection problem is a discrete problem, a modification of the initially proposed Artificial Bee Colony optimization algorithm, a Discrete Artificial Bee Colony optimization algorithm, is proposed in this study. The performance of the algorithm is compared with other popular metaheuristic methods like classic genetic algorithms, tabu search, GRASP, ant colony optimization, particle swarm optimization and honey bees mating optimization algorithm. In order to assess the efficacy of the proposed algorithm, this methodology is evaluated on datasets from the UCI Machine Learning Repository. The high performance of the proposed algorithm is achieved as the algorithm gives very good results and in some instances the percentage of the corrected clustered samples is very high and is larger than 98%.
Keywords: Artificial Bee Colony, Greedy Randomized Adaptive Search Procedure, Clustering Analysis.

1. INTRODUCTION

Artificial Bee Colony (ABC) optimization algorithm is a population-based swarm intelligence algorithm that was originally proposed by [14, 15] and it simulates the foraging behaviour that a swarm of bees perform. In this algorithm there are three groups of bees, the employed bees (bees that determines the food source (possible solutions) from a prespecified set of food sources and share this information (waggle dance) with the other bees in the hive), the onlookers bees (bees that based on the information that they take from the employed bees they search for a better food source in the neighborhood of the memorized food sources) and the scout bees (employed bees that their food source has been abandoned and they search for a new food source randomly). The initially proposed Artificial Bee Colony optimization algorithm is applied in continuous optimization problems. In our study, as the feature selection problem is a discrete problem, we made some modifications in the initially proposed algorithm in order to be suitable for solving these kind of problems.

In this paper, a new hybrid nature inspired intelligent technique, that uses the Discrete Artificial Bee Colony optimization algorithm (DABC) and the Greedy Randomized Adaptive Search Procedure (GRASP) [9] is presented and analyzed in detail for the solution of the clustering problem. More precisely, the proposed DABC-GRASP algorithm uses the DABC for the feature selection phase of the clustering algorithm while for the clustering phase the GRASP algorithm is applied. In order to assess the efficacy of the proposed algorithm, this methodology is evaluated on datasets from the UCI Machine Learning Repository. Also, the method is compared with the results of a number of other metaheuristic algorithms for clustering analysis that use, mainly, hybridization techniques that incorporate a Tabu Search based algorithm [10], a Genetic based algorithm [11], a Particle Swarm based Optimization algorithm [16], a Honey Bees Mating based Optization algorithm [1, 2] and an Ant Colony based Optimization algorithm [7]. The rest of this paper is organized as follows: In the next section a short literature review of the bees inspired optimization algorithms is presented, in section 3 the proposed Hybrid DABC-GRASP algorithm is presented and analyzed in detail. In section 4, the analytical computational results for the datasets used in this study are presented while in the last section conclusions and future research are given.

2. BEES INSPIRED OPTIMIZATION ALGORITHMS

During the last decade, nature inspired intelligence has become increasingly popular through the development and utilization of intelligent paradigms in advanced information systems design. When the task is optimization within complex domains of data or information, the most popular nature inspired approaches are those representing successful animal and micro-organism team behaviour, such as swarm or flocking intelligence (birds flocks or fish schools inspired Particle Swarm Optimization [16]), artificial immune systems [5, 6] (that mimic the biological one), ant colonies (ants foraging behaviors gave rise to Ant Colony Optimization [7]), or optimized performance of bees.

A number of swarm intelligence algorithms, based on the behaviour of the bees have been presented ([4]). These algorithms are divided, mainly, in two categories according to their behaviour in the nature, the foraging behaviour and the mating behaviour. The most important approaches that simulate the foraging behaviour of the bees are the Artificial Bee Colony (ABC) Algorithm proposed by [14, 15], the Virtual Bee Algorithm proposed by [30], the Bee

978-1-4244-4136-5/09/$ 25.00 ©2009 IEEE

Colony Optimization Algorithm proposed by [27], the BeeHive algorithm proposed by [28], the Bee Swarm Optimization Algorithm proposed by [8] and the Bees Algorithm proposed by [24]. The Virtual Bee Algorithm ([30]) is applied in continuous optimization problems. In this algorithm, the population of the bees are associated with a memory, a food source, and then all the memories communicate between them with a waggle dance procedure. The whole procedure is similar with a genetic algorithm and it has been applied on two function optimization problems with two parameters. In the BeeHive ([28]) algorithm, a protocol inspired from dance language and foraging behaviour of honey bees is used. In the Bees Swarm Optimization ([8]), initially a bee finds an initial solution (food source) and from this solution the other solutions are produced with certain strategies. Then, every bee is assigned in a solution and when they accomplish their search, the bees communicate between them with a waggle dance strategy and the best solution will become the new reference solution. To avoid cycling the authors use a tabu list. In the Bees Algorithm ([24]), a population of initial solutions (food sources) are randomly generated. Then, the bees are assigned to the solutions based on their fitness function. The bees return to the hive and based on their food sources, a number of bees are assigned to the same food source in order to find a better neighborhood solution. In the Bee Colony Optimization ([27]) algorithm, a step by step solution is produced by each forager bee and when the foragers return to the hive a waggle dance is performed by each forager. Then the other bees, based on a probability, follow the foragers. This algorithm looks like the Ant Colony Optimization ([7]) algorithm but it does not use at all the concept of pheromone trails.

Contrary to the fact that there are many algorithms that are based on the foraging behaviour of the bees, the main algorithm proposed based on the marriage behaviour is the Honey Bees Mating Optimization Algorithm (HBMO), that was presented in ([1, 2]). The Honey Bees Mating Optimization algorithm simulates the mating process of the queen of the hive. The mating process of the queen begins when the queen flights away from the nest performing the mating flight during which the drones follow the queen and mate with her in the air ([1, 3]). The algorithm is a swarm intelligence algorithm since it uses a swarm of bees where there are three kinds of bees, the queen, the drones and the workers. There is a number of procedures that can be applied inside the swarm. In the honey bees mating optimization algorithm, the procedure of mating of the queen with the drones is described. First, the queen is flying randomly in the air and, based on her speed and her energy, if she meets a drone then there is a possibility to mate with him. Even if the queen mates with the drone, she does not create directly a brood but stores the genotype (with the term "genotype" we mean some of the basic characteristics of the drones, i.e. part of the solution) of the drone in her spermatheca and the brood is created only when the mating flight has been completed. A crossover operator is used in order to create the broods. In a hive the role of the workers is simply the brood care (i.e. to feed them with the "royal jelly") and, thus, they are only a local search phase in the honey bees mating optimization algorithm. And thus, this algorithm combines both the mating process of the queen and one part of the foraging behavior of the honey bees inside the hive. If a broods is better (more fittest) than the queen, then this brood replaces the queen.

3. THE PROPOSED HYBRID DABC-GRASP FOR CLUSTERING

3.1. Clustering Problem

Clustering analysis identifies clusters (groups) embedded in the data, where each cluster consists of objects that are similar to one another and dissimilar to objects in other clusters ([13, 25, 29]). The typical cluster analysis consists of four steps (with a feedback pathway) which are the ***feature selection***, that is an optimization problem, where the problem is to search through the space of feature subsets to identify the optimal or near-optimal one with respect to a performance measure (see [12] for feature selection algorithms) or ***feature extraction*** where some transformations are used in order to generate useful and novel features from the original ones, the ***clustering algorithm design*** or ***selection*** that is usually combined with the selection of a corresponding proximity measure ([13, 25]) and with the construction of a clustering criterion function which makes the partition of clusters a well defined optimization problem (see ([13, 25, 29]) for an analytical survey of the clustering algorithms), the ***cluster validation*** where external indices, internal indices, and relative indices are used for cluster validity analysis ([13, 29]) and the ***results interpretation*** where experts in the relevant fields interpret the data partition in order to guarantee the reliability of the extracted knowledge ([29]).

More precisely, the problem of clustering N objects (patterns) into K clusters is considered. In particular the problem is stated as follows: Given N objects in R^n, allocate each object to one of K clusters such that the sum of squared Euclidean distances between each object and the center of its belonging cluster (which is also to be found) for every such allocated object is minimized. For the mathematical description of the clustering problem see ([17, 19, 18, 21, 23, 22].)

The proposed algorithm (Hybrid DABC-GRASP) for the solution of the clustering problem is a two phase algorithm which combines a Discrete Artificial Bee Colony Optimization (DABC) algorithm for the solution of the feature selection problem and a Greedy Randomized Adaptive Search Procedure (GRASP) for the solution of the clustering problem. In this algorithm, the activated features are calculated by the DABC (see 3.2) and the fitness (quality) of each food source is calculated by the clustering algorithm (see 3.3). The clustering algorithm has the possibility to solve the clustering problem with known or unknown number of clusters (see 3.3 and [17, 19, 18, 21, 23, 22]).

3.2. DABC for the Feature Selection Problem

Feature selection is used as the first step of the clustering task in order to reduce the dimension of problem, decrease noise and improve the speed of the algorithm by the elimination of irrelevant or redundant features. In this paper, an extended version of the Artificial Bee Colony optimization algorithm [14, 15] in the discrete space, the DABC, is proposed for the feature selection.

In the Artificial Bee Colony optimization algorithm [14, 15], there are three kind of artificial bees in the colony, the employed bees, the onlooker bees and the scouts. Initially, a set of food source positions (possible solutions) are randomly selected by the employed bees and their nectar amounts (fitness functions) are determined. One of the key issues in designing a successful algorithm for Feature Selection Problem is to find a suitable mapping between Feature Selection Problem solutions and food sources in Artificial Bee Colony Optimization Algorithm. In the Artificial Bee Colony optimization algorithm [14, 15], the food sources are randomly generated. However in this study, the solutions should have values equal to 0 or to 1, where 0 denotes that the feature is not activated and 1 denotes that the feature is activated. In our proposed algorithm, the food sources are calculated exactly as in the initially proposed algorithm and, then, the values are transformed by using a sigmoid function:

$$sig(x_{ij}) = \frac{1}{1 + exp(-x_{ij})} \quad (1)$$

and then the food sources are calculated by:

$$y_{ij} = \begin{cases} 1, & \text{if } rand1 < sig(x_{ij}) \\ 0, & \text{if } rand1 >= sig(x_{ij}) \end{cases} \quad (2)$$

where x_{ij} is the solution (food source), $i = 1, ..., N$ (N is the number of food sources), $j = 1, ..., d$ (d is the dimension of the problem - the number of features), y_{ij} is the transformed solution and $rand1$ is a random number in the interval (0,1). These equations have also been used for the Discrete Particle Swarm Optimization [26]. Afterwards, the fitness of each food source is calculated using the GRASP algorithm for clustering (see section 3.3) and an employed bee is attached to each food source. The employed bees return in the hive and perform the waggle dance in order to inform the other bees (onlooker bees) about the food sources. Then, the onlooker bees choose the food source that will visit based on the nectar information taken from the waggle dance of the employed bees. The probability of choosing a food source is given by [14, 15]:

$$p_i = \frac{f_i}{\sum_{n=1}^{N} f_n} \quad (3)$$

where f_i is the fitness function of each food source. As it was mentioned previously the nectar information corresponds to the fitness function of each food source. Since the clustering is a minimization problem, if a solution has a high value in the cost function, then it is not a good solution for the clustering problem and its fitness value must be small. So a high fitness value must correspond to a solution with a small cost function. A way to accomplish this is to find initially the solution in the population with the maximum cost and to subtract from this value the cost of each of the other solutions. Now, the higher fitness value corresponds to the solution with the shorter cost. Since the probability of selecting a food source by the onlooker bees is related to its fitness, and since the food source with the worst cost has fitness equal to zero, it will never be selected for food gathering. To avoid this possibility the fitness of all food sources is incremented by one.

Afterwards, the employed and the onlooker bees are placed in the selected food sources. In order to produce a new food position from the old one the Discrete Artificial Bee Colony algorithm, uses the same equation as in the Artificial Bee Colony Algorithm [14, 15]:

$$x'_{ij} = x_{ij} + rand2(x_{ij} - x_{kj}) \quad (4)$$

where x'_{ij} is the candidate food source and k is a different from i food source and $rand2$ is a random number in the interval (0,1). As the values of the candidate food source are not suitable for the clustering problem they are transformed to the y'_{ij} using the equations (1) and (2). Afterwards, the fitness of each food source is calculated using the GRASP algorithm for clustering. It should be noted that if there is a large number of bees in a food source, then from the local search moves that each bee performs, this food source has larger exploration abilities in each iteration. If a better food source is found in an iteration, this food source replaces the old one. If for a number of iterations a solution is not improved, then this solution is assumed to be abandoned and a scouter bee is placed in a new random position (a new food source).

3.3. GRASP for the Clustering Problem

As it was mentioned earlier in the clustering phase of the proposed algorithm a **Greedy Randomized Adaptive Search Procedure (GRASP)** ([9]) is used which is an iterative two phase search algorithm (a **construction phase** and a **local search phase**). An initial solution (i.e. an initial clustering of the samples in the clusters) is constructed step by step and, then, this solution is exposed for improvement in the local search phase of the algorithm. The first problem that we had to face was the selection of the number of the clusters. Thus, the algorithm works with two different ways.

If the number of clusters is known a priori, then a number of samples equal to the number of clusters are selected randomly as the initial clusters. In this case, as the iterations of GRASP increase, the number of clusters remains unchanged. In each iteration, different samples (equal to the number of clusters) are selected as initial clusters. Afterwards, the RCL is created (the Restricted Candidate List - RCL is the list that is used for the selection of the next element that will be chosen to be inserted to the current solution). The probabilistic component of a **GRASP** is characterized by randomly choosing one of the best candidates in the list but not necessarily the top candidate. In our implementation, the best promising candidate samples are selected to create the RCL. The samples in the list are ordered taking into account the distance of each sample from all centers of the clusters and the ordering is from the smallest to the largest distance. From this list, the first D samples (D is a parameter of the problem) are selected in order to form the final RCL. The candidate sample for inclusion in the solution is selected randomly from the RCL using a random number generator. Finally, the RCL is readjusted in every iteration by recalculating all the distances based on the new centers and replacing the sample which has been included in the solution by another sample that does not belong to the RCL, namely the $(D + t_1)$th sample where t_1 is the number of the current iteration. When all the samples have been assigned to clusters a local search strategy is applied in order to improve the solution. The local search works as follows: For each sample the probability of its reassignment in a different cluster is examined by calculating the distance of the sample from the centers. If a sample is reassigned to a different cluster the new centers are calculated. The local search phase stops when in an iteration no sample is reassigned. If the number of clusters is unknown then, initially a number of samples are selected randomly as the initial clusters. Now, as the iterations of GRASP increase, the number of clusters changes but cannot become less than two. In each iteration a different number of clusters can be found. The creation of the initial solutions and the local search phase work as in the previous case. The only difference compared to the previous case concerns the use of the validity measure in order to choose the best solution ([17, 18, 19, 20, 21, 22, 23]).

4. Computational Results

The performance of the proposed methodology is tested on 9 benchmark instances taken from the UCI Machine Learning Repository [31]. The datasets from the UCI Machine Learning Repository were chosen to include a wide range of domains and their characteristics are given in Table 1. The data varies in terms of the number of observations from very small samples (Iris with 150 observations) up to larger data sets (Spambase with 4601 observations). Also, there are data sets with two and three clusters. In one case (Breast Cancer Wisconsin) the data set is appeared with different size of observations because in this data set there is a number of missing values. The problem of missing values was faced with two different ways. In the first way where all the observations are used we took the mean values of all the observations in the corresponding feature while in the second way where we have less values in the observations we did not take into account the observations that they had missing values. Some data sets involve only numerical features, and the remaining include both numerical and categorical features (in parentheses in Table 1). For each data set, Table 1 reports the total number of features and the number of categorical features in parentheses. The parameter settings for DABC-GRASP algorithm were selected after thorough empirical testing and they are: The number of employed is set equal to 50, the number of onlookers is set equal to 100, the number of generations is set equal to 100, the size of RCL is set equal to 50, the number of GRASP's iterations is equal to 100. The algorithm was implemented in Fortran 90 and was compiled using the Lahey f95 compiler on a Centrino Mobile Intel Pentium M 750 at 1.86 GHz, running Suse Linux 9.1.

The objective of the computational experiments is to show the performance of the proposed algorithm in searching for a reduced set of features with high clustering of the data. The purpose of feature variable selection is to find the smallest set of features that can result in satisfactory predictive performance. Because of the curse of dimensionality, it is often necessary and beneficial to limit the number of input features in order to have a good predictive and less computationally intensive model. In general there are $2^{NF} - 1$ possible feature combinations, where NF denotes the number of features. Thus, in our cases the problem with the fewest number of feature combinations is the Iris (namely $2^4 - 1$), while the most difficult problem is the Spambase where the number of feature combinations is $2^{57} - 1$.

Tab. 1: Data Sets Characteristics

Data Sets	Observations	Features	Clusters
Australian Credit (AC)	690	14(8)	2
Breast Cancer Wisconsin 1 (BCW1)	699	9	2
Breast Cancer Wisconsin 2 (BCW2)	683	9	2
Heart Disease (HD)	270	13(7)	2
Hepatitis 1 (Hep1)	155	19 (13)	2
Ionosphere (Ion)	351	34	2
Spambase(spam)	4601	57	2
Iris	150	4	3
Wine	178	13	3

A comparison with the classic k-means and other metaheuristic approaches for the solution of the clustering problem is presented in Table 2. In this Table, besides the proposed algorithm, eleven other algorithms are used for the solution of the feature subset selection problem and for the clustering problem.

In the first group of algorithms, besides the proposed algorithm, the results of three other algorithms are presented where in the feature selection phase a Multi Swarm Constriction Particle Swarm Optimization (MSCPSO), a Memetic algorithm and a Honey Bees Mating Optimization (HBMO) are used, respectively, while in the clustering phase a GRASP algorithm is used. In the second group of algorithms, all algorithms use the GRASP algorithm in the clustering phase while in the feature selection phase a PSO algorithm, an ACO algorithm, a genetic algorithm and a tabu search algorithm are used, respectively. In the third group of algorithms, initially the classic k-means algorithm is used for the clustering problem using all features, while in the rest columns an ACO algorithm is used in both phases and, finally, a PSO algorithm is used in both phases. The other algorithm of the last group uses a PSO algorithm in the feature selection phase and an ACO algorithm in the clustering phase. The parameters and the implementation details of all of the algorithms presented in the comparisons are analyzed in the papers [17, 18, 19, 20, 21, 22, 23].

From this table, it can be observed that the Hybrid DABC-GRASP algorithm performs better (has the largest number of correct clustered samples) than the other eleven algorithms in all instances. It should be mentioned that in some instances the differences in the results between the Hybrid DABC-GRASP algorithm and the other ten algorithms are very significant. Mainly, for the two data sets that have the largest number of features compared to the other data sets, i.e. in the Ionosphere data set the percentage of correct clustered samples for the Hybrid DABC-GRASP algorithm is 90.02% while for all the other methods the percentage varies between 70.65% to 88.88% and in the Spambase data set the percentage of correct clustered samples for the Hybrid DABC-GRASP algorithm is 90.65% while for all the other methods the percentage varies between 82.80% to 90.17%. It should, also, be noted that a hybridization

Tab. 2: Results of the algorithms

Instance	DABC-GRASP		MSCPSO-GRASP		Memetic-GRASP		HBMO-GRASP	
	Selected Feat.	Correct Clustered	Selected Features	Correct Clustered	Selected Feat.	Correct Clustered	Selected Feat.	Correct Clustered
BCW2	5	669(97.95%)	5	667(97.65%)	5	664(97.21%)	5	664(97.21%)
Hep1	5	144(92.90%)	5	142(91.65%)	9	139(89.67%)	5	140(90.32%)
AC	7	612(88.69%)	7	610(88.40%)	8	604(87.53%)	8	604(87.53%)
BCW1	4	682(97.56%)	5	681(97.42%)	8	677(96.85%)	5	677(96.85%)
Ion	5	316(90.02%)	5	312(88.88%)	5	305(86.89%)	8	309 (88.03%)
spam	24	4171(90.65%)	28	4149(90.17%)	32	4019(87.35%)	31	4028 (87.54%)
HD	5	248(91.85%)	6	243(90.00%)	9	236(87.41%)	8	237(87.77%)
Iris	3	147(98.00%)	3	147(98.00%)	3	146(97.33%)	3	146(97.33%)
Wine	6	177(99.43%)	6	177(99.43%)	7	176(98.87%)	7	176(98.87%)
Instance	PSO-GRASP		ACO-GRASP		Genetic-GRASP		Tabu-GRASP	
	Selected Features	Correct Clustered	Selected Feat.	Correct Clustered	Sel. Feat.	Correct Clustered	Sel. Feat.	Correct Clustered
BCW2	5	662(96.92%)	5	662(96.92%)	5	662(96.92%)	6	661(96.77%)
Hep1	7	135(87.09%)	9	134(86.45%)	9	134(86.45%)	10	132(85.16%)
AC	8	604(87.53%)	8	603(87.39%)	8	602(87.24%)	9	599(86.81%)
BCW1	5	676(96.70%)	5	676(96.70%)	5	676(96.70%)	8	674(96.42%)
Ion	11	300(85.47%)	2	291(82.90%)	17	266(75.78%)	4	263(74.92%)
spam	51	4009(87.13%)	56	3993(86.78%)	56	3938(85.59%)	34	3810(82.80%)
HD	9	232(85.92%)	9	232(85.92%)	7	231(85.55%)	9	227(84.07%)
Iris	3	145(96.67%)	3	145(96.67%)	4	145(96.67%)	3	145(96.67%)
Wine	7	176(98.87%)	8	176(98.87%)	7	175(98.31%)	7	174(97.75%)
Instance	k-Means		ACO		PSO		PSO-ACO	
	Sel Feat.	Correct Clustered	Selected Features	Correct Clustered	Sel. Feat.	Correct Clustered	Selected Features	Correct Clustered
BCW2	9	654(95.74%)	5	662(96.92%)	5	662(96.92%)	5	664(97.21%)
Hep1	19	121(78.06%)	9	133(85.80%)	10	132(85.16%)	6	139(89.67%)
AC	14	580(84.05%)	8	601(87.10%)	8	602(87.24%)	8	604(87.53%)
BCW1	9	672(96.13%)	8	674(96.42%)	8	674(96.42%)	5	677(96.85%)
Ion	34	248(70.65%)	16	258(73.50%)	12	261(74.35%)	7	302(86.03%)
spam	57	3958(86.02%)	41	3967(86.22%)	37	3960(86.06%)	39	4012(87.19%)
HD	13	220(81.48%)	9	227(84.07%)	9	227(84.07%)	9	235(87.03%)
Iris	4	144(96%)	3	145(96.67%)	3	145(96.67%)	3	146(97.33%)
Wine	13	172(96.92%)	7	174(97.75%)	7	174(97.75%)	7	176(98.87%)

algorithm performs always better than a no hybridized algorithm. More precisely, the only six algorithms that are competitive in almost all instances with the proposed Hybrid DABC-GRASP algorithm are the Hybrid MSCPSO-GRASP, the Hybrid HBMO-GRASP, the Hybrid Memetic-GRASP, the Hybrid PSO - ACO, the Hybrid PSO - GRASP and the Hybrid ACO - GRASP algorithms. These results prove the significance of the solution of the feature selection problem in the clustering algorithm as when more sophisticated methods for the solution of this problem (Artificial Bee Colony, Particle Swarm Optimization, Honey Bees Mating Optimization, and Ant Colony Optimization) were used the performance of the clustering algorithm was improved. The significance of the solution of the feature selection problem using the Discrete Artificial Bee Colony Algorithm is, also, demonstrated by the fact that with this algorithm the best solution was found by using less features than the other algorithms used in the comparisons. More precisely, in the most difficult instance, the Spambase instance, the proposed algorithm needed 24 features in order to find the optimal solution, while the other ten algorithms (in the k-means the feature selection problem was not solved) the algorithms needed between 28 - 56 features to find their best solution. It should, also, be mentioned that the algorithm was tested with two options: with known and unknown number of clusters. When the number of clusters was unknown and, thus, in each iteration of the algorithm different initial values of clusters were selected, the algorithm always converged to the optimal number of clusters and with the same results as in the case where the number of clusters was known.

5. Conclusions and Future Research

In this paper, a new metaheuristic algorithm, the Hybrid DABC-GRASP, is proposed for solving the Clustering Problem. This algorithm is a two phase algorithm which combines a Discrete Artificial Bee Colony Optimization algorithm for the solution of the feature selection problem and a Greedy Randomized Adaptive Search Procedure for the solution of the clustering problem. A number of other metaheuristic algorithms for the solution of the problem were also used for comparison purposes. The performance of the proposed algorithm was tested using various benchmark datasets from UCI Machine Learning Repository. The significance of the solution of the clustering problem by the proposed algorithm is demonstrated by the fact that the percentage of the correct clustered samples is very high and in some instances is larger than 98%. Future research is intended to be focused in using different algorithms both to the feature selection phase and to the clustering algorithm phase.

REFERENCES

[1] H.A. Abbass, "A monogenous MBO approach to satisfiability", *Proceeding of the International Conference on Computational, Intelligence for Modelling, Control and Automation, CIMCA'2001. Las Vegas, NV, USA*, 2001.

[2] H.A. Abbass, "Marriage in honey-bee optimization (MBO): a haplometrosis polygynous swarming approach", *The Congress on Evolutionary Computation (CEC2001), Seoul, Korea, May 2001*, pp. 207-214, 2001.

[3] Afshar, A., Bozog Haddad, O., Marino, M.A., B.J. Adams, "Honey-bee mating optimization (HBMO) algorithm for optimal reservoir operation, *Journal of the Franklin Institute*, Vol. 344, pp. 452-462, 2007.

[4] Baykasoglu, A., Ozbakor, L., P. Tapkan, "Artificial bee colony algorithm and its application to generalized assignment problem", *Swarm Intelligence, Focus on Ant and Particle Swarm Optimization, I-Tech Education and Publishing, Chan, F T S, Tiwari M K (Eds)*, pp. 113 - 144, 2007.

[5] D. Dasgupta (Ed.), "Artificial immune systems and their application", *Springer, Heidelberg*, 1998.

[6] De Castro, L.D., J. Timmis, "Artificial immune systems: A new computational intelligence approach", *Springer, Heidelberg*, 2002.

[7] Dorigo, M., T. Stutzle, "Ant colony optimization", *A Bradford Book, The MIT Press Cambridge, Massachusetts, London, England*, 2004.

[8] Drias, H., Sadeg, S., S. Yahi, "Cooperative bees swarm for solving the maximum weighted satisfiability problem", *IWAAN International Work Conference on Artificial and Natural Neural Networks, LNCS*, Vol. 3512, pp. 318-325, 2005.

[9] Feo, T.A., M.G.C. Resende, "Greedy randomized adaptive search procedure", *Journal of Global Optimization*, Vol. 6, pp. 109-133, 1995.

[10] F. Glover, "Tabu search I", *ORSA Journal on Computing*, Vol. 1, No. 3, pp. 190-206, 1989.

[11] D.E. Goldberg, "Genetic algorithms in search, optimization, and machine learning", *Addison-Wesley Publishing Company, INC, Massachussets*, 1989.

[12] Jain, A., D. Zongker, "Feature selection: Evaluation, application, and small sample performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 153-158, 1997.

[13] Jain, A.K., Murty, M.N., P.J. Flynn, "Data clustering: A review", *ACM Computing Surveys*, Vol. 31, No. 3, pp. 264-323, 1999.

[14] Karaboga, D., B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm", *Journal of Global Optimization*, DOI 10.1007/s10898-007-9149-x, 2007.

[15] Karaboga, D., B. Basturk, "On the performance of artificial bee colony (ABC) algorithm", *Applied Soft Computing*, Vol. 8, pp. 687-697, 2008.

[16] Kennedy, J., R. Eberhart, "Particle swarm optimization", *Proceedings of 1995 IEEE International Conference on Neural Networks*, Vol. 4, pp. 1942-1948, 1995.

[17] Marinakis, Y., Marinaki, M., N. Matsatsinis, "A hybrid particle swarm optimization algorithm for cluster analysis", *Data Warehouse and Knowledge Discovery 2007, LNCS 4654, I.S. Song, J. Eder and T.M. Nguyen (Eds.), Springer Berlin Heidelberg New York*, pp. 241-250, 2007.

[18] Marinakis, Y., Marinaki, M., N. Matsatsinis "A hybrid clustering algorithm based on honey bees mating optimization and greedy randomized adaptive search procedure", *Learning and Intelligence Optimization - LION 2007, LNCS, Springer-Verlag Berlin Heidelberg*, Vol. 5313, pp. 138-152, 2008.

[19] Marinakis, Y., Marinaki, M., N. Matsatsinis, "A stochastic nature inspired metaheuristic for clustering analysis", *International Journal of Business Intelligence and Clustering Analysis*, Vol. 3, No. 1, pp. 30 - 44, 2008.

[20] Marinakis, Y., Marinaki, M., N. Matsatsinis, "A hybrid clustering algorithm based on multi-swarm constriction PSO and GRASP", *Data Warehouse and Knowledge Discovery 2008, LNCS, I.S. Song, J. Eder and T.M. Nguyen (Eds.), Springer Berlin Heidelberg New York*, Vol. 5182, pp. 186-195, 2008.

[21] Marinakis, Y., Marinaki, M., Matsatsinis, N., C. Zopounidis, "A memetic-GRASP algorithm for clustering", *10th International Conference on Enterprise Informations Systems, 13 -16 June 2008, Barcelona, Spain*, 2008.

[22] Marinakis, Y., Marinaki, M., Doumpos, M., Matsatsinis, N., C. Zopounidis, "A hybrid stochastic genetic - GRASP algorithm for clustering analysis", *Operational Research: An International Journal*, Vol. 8, No. 1, pp. 33-46, 2008.

[23] Marinakis, Y., Marinaki, M., Doumpos, M., Matsatsinis, N., C. Zopounidis, "A hybrid ACO-GRASP algorithm for clustering analysis", *Annals of Operations Research*, (accepted), 2009.

[24] Pham D.T., Kog, E., Ghanbarzadeh, A., Otri, S., Rahim, S., M. Zaidi, "The bees algorithm - A novel tool for complex optimization problems", *IPROMS 2006 Proceeding 2nd International Virtual Conference on Intelligent Production Machines and Systems, Oxford, Elsevier*, 2006.

[25] Rokach, L., O. Maimon "Clustering methods", *Data Mining and Knowledge Discovery Handbook, Maimon, O. and L. Rokach (Eds.), Springer, New York*, pp. 321-352, 2005.

[26] Shi, Y., R. Eberhart, "A modified particle swarm optimizer", *Proceedings of 1998 IEEE World Congress on Computational Intelligence*, pp. 69-73, 1998.

[27] Teodorovic, D., M. Dell'Orco, "Bee colony optimization - A cooperative learning approach to complex transportation problems", *Advanced OR and AI Methods in Transportation*, pp. 51-60, 2005.

[28] Wedde, H.F., M. Farooq, Y. Zhang, "BeeHive: An efficient fault-tolerant routing algorithm inspired by honey bee behavior". *Ant colony optimization and swarm intelligence, Dorigo, M., (Ed.), LNCS, Springer Berlin*, Vol. 3172, pp, 83-94, 2004.

[29] Xu, R., D. Wunsch II, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, pp. 645-678, 2005.

[30] X.S. Yang, "Engineering optimizations via nature-inspired virtual bee algorithms" *IWINAC 2005, Yang, J. M., Alvarez, J.R., (Eds.), LNCS, Springer-Verlag, Berlin Heidelberg*, Vol. 3562, pp. 317-323, 2005.

[31] http://archive.ics.uci.edu/ml/