# A HYBRID APPROACH FOR CLUSTERING BASED ON COMBINATIONAL ALGORITHMS

Mrs.S.Vidhiyashri
Student, Department of Computer science and Engineering,
SSM College OF Engineering, Komarapalayam.
svidhiyashri@gmail.com

**Abstract - Data clustering helps one discern the structure of and simplify the complexity of massive quantities of data. It is a common technique for statistical data analysis and is used in many fields, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics. The well-known K-means algorithm, which has been successfully applied to many practical clustering problems, suffers from several drawbacks due to its choice of initializations. However, its performance depends on the initial state of centroids and may trap in local optima. The gravitational search algorithm (GSA) is one effective method for find optimal solution. The GSA-KM algorithm helps the k means algorithm to escape from local optima and also increases the convergence speed of the GSA algorithm. A hybrid technique based on combining the K-means algorithm, Gravitational Search algorithm, Nelder–Mead simplex search, and particle swarm optimization, called KM–GSA-NM–PSO, is proposed. The KM-GSA–NM–PSO searches for cluster centers of an arbitrary data set as does the K-means algorithm, but it can effectively and efficiently find the global optima. The new KM–GSA-NM–PSO algorithm is tested on UCI repository data sets, and its performance is compared with those of K means and KM-GSA clustering algorithms. Enhancement can be made to this algorithm such as image segmentation and university time tabling.**

**Keywords:** Data clustering; K-means clustering; Nelder–Mead simplex search method; Particle swarm optimization, GSA algorithm

## I. INTRODUCTION

Clustering is a time consuming and tedious activity in the field of data mining. The data that is acquired by means of clustering is manifold varied and complicated. So, consequently the results obtained from Clustering may not be as objective and precise. Also, the various methods of clustering which are available may provide results with variable results of accuracy as each follow their own method and algorithm to generate the clusters. Some clustering algorithms may not provide good enough results (Gravitational Search Algorithm), while some clustering algorithms, though they work very well providing fairly good results are bound by a constraint/condition that needs to be fulfilled and satisfied for their accurate and successful working (Initial value of "K" is needed for K-Means Clustering Algorithm). Particle Sparm Optimization (PSO), a population based algorithm has a slow convergence rate. This problem can be solved by using the Nelder Mead (NM), a local line search method.

The motivation for the undertaking of this project is in the aim to capture and provide far more accurate results for clustering, in a more efficient manner than those provided by the clustering algorithms available today.

Clustering is a process where objects with similar characteristics are grouped together. It is a method of unsupervised learning where any initial knowledge of the dataset is not necessary. Clustering is of many types and varied in its operation. The basic variations of clustering are Hierarchical Clustering and Partitional Clustering.

There are many algorithms that have been proposed to perform clustering. However, due to a large variety of applications, different data types and various purposes of clustering, we cannot find a unique algorithm that can serve all the requirements at once. In general, clustering algorithms can be divided into two groups: hierarchical algorithms and partitioned algorithms. Hierarchical clustering algorithms recursively find clusters either in an agglomerative (bottom–up) mode or in a divisive (top–down) mode. Agglomerative methods start with each data object in a separate cluster and successively merge the most similar pairs until termination criteria are satisfied. Divisive methods start with all the data objects in one cluster and repeatedly divide each cluster into smaller clusters, also until termination criteria are satisfied. On the other hand, partitional clustering algorithms find all the clusters simultaneously without forming a hierarchical structure. A well-known class of partitional clustering algorithms is the centre-based clustering method, and the most popular widely used algorithm from this class of algorithms is a k-means algorithm. K-means is simple to implement and efficient in most cases [16–19].However, the performance of k-means is highly dependent on the initial state of centroids and may converge to the local optima rather than global optima. The k-means algorithm tries to minimize the intra-cluste rvariance; butit does not ensure that the result has a global minimum variance [20, 21].

The Objective of this project to improve the results and quality of clustering by combining the various approaches eliminating the drawbacks of the individual algorithm, to reduce the Intra cluster distance between the datasets by hybrid algorithms, to increase the efficiency and accuracy of the clusters and to extend the results of clustering to other fields of datasets and to check their viability and effectiveness

## II. BACKGROUND ON CLUSTERING ALGORITHMS TO BE USED

### A. K-MEANS ALGORITHM IMPLEMENTATION.

K-Means Method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed. At this point we need to re-calculate k new centroids as bary center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change. Their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is

$$J = \sum_{j=1}^{k} \sum_{i=1}^{k} \left\| x_i^{(j)} \quad - \quad c_j \right\|^2 \tag{1}$$

Where $\left\| x_i^{(j)} \quad - \quad c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre $c_j$, is an indicator of the distance of the *n* data points from their respective cluster centers.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

### B. GSA ALGORITHM IMPLEMENTATION.

Gravitational search algorithm (GSA) based on the law of gravity and the notion of mass interactions. It is defined by Newton as, ‘‘Every particle in the universe attracts every other particle with a force that is directly proportional to the product of the masses of the particles and inversely proportional to the square of the distance between them’’

It is formulated as

$$F = G\ (M_1 \text{ x } M_2)\ /\ R^2 \tag{2}$$

F is the gravitational force
G is the gravitational constant (value of $6.67259 \text{ x } 10^{-11}$)
$M_1$and $M_2$ are the masses of first and second particles
R is the straight-line distance between the two particles

In GSA, there is an isolated system of masses. Using the gravitational force, every mass in the system can see the situation of other masses. In GSA, agents are considered as objects and their performance is measured by their masses. All these objects attract each other by a gravity force. This force causes a movement of all objects globally towards the objects with heavier masses. The heavy masses correspond to good solutions of the problem. The position of the agent corresponds to a solution of the problem, and its mass is determined using a fitness function.

### C. NELDER-MEADALGORITHM IMPLEMENTATION.

NM simplex search method proceeds by evaluating the fitness function values at the (N+1) vertices of an initial simplex.The highest fitness function value will be replaced by a newly reflected and better point, which can be located in the negative gradient direction.NM is a direct line search method of the steepest descent kind. The ingredients of the replacement process consist of four basic operations: reflection, expansion, contraction, and shrinkage.An example of minimization of a function of two variables (N = 2) will illustrate the basic procedure of NM. Starting with point B together with an initial step size, an initial simplex design shown as A, B and C is constructed, as illustrated in Figure1

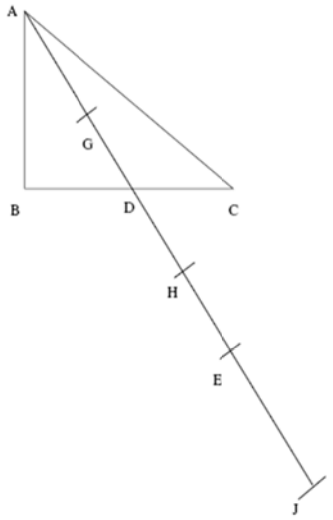

Figure.1. NM operations on a two-dimensional case.

1. Sort the function values at A, B, and C. Supposef(C) < f(B) < f(A). f(A) is the highest of the three function values and is to be replaced. In this case, a reflection is made through the centroid of BC at point D to point E.

2. If f(E) < f(C), an expansion is made to point J. We then keep E or J as a replacement for A, depending on which function value is lower.

3. If f(E) > f(C), a contraction is made to point G or H as a replacement for A, depending on which of f(A) and f(E) is lower, provided that f(G) or f(H) is smaller than f(C).

4. If either f(G) or f(H) is larger than f(C), the contraction has failed and we then perform a shrinkage operation. The shrinkage operation reduces the size of the simplex by moving all but the best point C halfway towards the best point C. We then have new points A and B. Go back to step 1.

**D. PSO algorithm implementation.**

PSO concept is based on a metaphor of social interaction such as bird flocking and fish schooling. Similar to genetic algorithms, PSO is also population-based and evolutionary in nature, with one major difference from genetic algorithms, which is that it does not implement filtering, i.e., all members in the population survive through the entire search process.

The steps of PSO are outlined below:

1. **Initialization**. Randomly generate 5N potential solutions, called ‘‘particles’’, N being the number of parameters to be optimized, and each particle is assigned a randomized velocity.

2. **Velocity update.** The particles then ‘‘fly’’ through hyperspace while updating their own velocity, which is accomplished by considering its own past flight and those of its companions’. The particle’s velocity and position are dynamically updated by the following equations:

$$V_{id}^{New} = w * V_{id}^{old} + c_1 * rand * \left(P_{id} - x_{id}^{old}\right) + c_2 * rand * \left(P_{id} + x_{id}^{old}\right) \quad (3)$$

$$X_{id}^{New} = X_{id}^{old} + V_{id}^{New} \quad (4)$$

where $c_1$ and $c_2$ are two positive constants, w is an inertia weight, and rand is a uniformly generated random number.

## III. EXISTING SYSTEM

The well-known K-means algorithm, which has been successfully applied to many practical clustering problems, suffers from several drawbacks due to its choice of initializations. A hybrid technique based on combining the K-means algorithm with various other algorithms is providing an improvement over the algorithm. Thus the combined approach of various algorithms provides a better performance using the goodness of the entire algorithm overcoming the disadvantage of any specific algorithm.

K-means is a simple and efficient algorithm that is widely used for data clustering. However, its performance depends on the initial state of centroids and may trap in local optima. The gravitational search algorithm (GSA) is one effective method for find optimal solution. Thus a hybrid data clustering algorithm based on GSA and k-means (GSA-KM), which uses the advantages of both algorithms .The GSA-KM algorithm helps the k means algorithm to escape from local optima and also increases the convergence speed of the GSA algorithm. We compared the performance of GSA-KM with other well-known algorithms, including k-means, genetic algorithm (GA), simulated annealing (SA), ant colony optimization (ACO), honey bee mating optimization (HBMO), particle swarm optimization (PSO) and gravitational search algorithm (GSA).

A hybrid technique based on combining the K-means algorithm, Nelder–Mead simplex search, and particle swarm optimization, called K–NM–PSO, is proposed in this research. The KM-GSA–NM–PSO searches for cluster centers of an arbitrary data set as does the K-means algorithm, but it can effectively and efficiently find the global optima. The new KM-GSA–NM–PSO algorithm is tested on various data sets, and its performance is compared with those of PSO, NM–PSO, K–PSO and K-means clustering. Results show that KM-GSA–NM–PSO is both robust and suitable for handling data clustering.

### A. Drawbacks

K-means Clustering is strongly dependent on initial representatives, A representative may be trapped in the local optimum during optimization, the presence of outliers influences clustering, It assumes that all attributes have equal importance for clustering and
The number of clusters is user-dependent.

In GSA algorithm, each agent could observe the performance of the others; the gravitational force is an information-transferring tool. Due to the force that acts on an agent from its neighborhood agents, it can see space around itself. A heavy mass has a large effective attraction radius and hence a great intensity of attraction. Therefore, agents with a higher performance have a greater gravitational mass. As a result, the agents tend to move toward the best agent.

Particle swarm optimization (PSO), a population-based algorithm which searches automatically for the optimum solution in the search space, and the searching process is not carried out at random. It is limited by the high computational cost of the slow convergence rate. The convergence rate of PSO is also typically slower than local search techniques.

## IV. PROPOSED SYSTEM

The above experimental results confirm that our existing hybrid approach has three significant merits in comparison to k-means and GSA alone. Firstly, it causes the k-means algorithm to escape from local optima. Secondly, it improves the quality of solutions found by either the k-means or GSA algorithm alone and thirdly it increases the convergence speed of the GSA algorithm.

It can be well understood that the hybrid algorithms includes all the best features of the existing algorithm overcoming the limitations of the individual algorithm when they are combined. Enhancing this combinational approach will leads to even better efficient results. This requires minimum number of function evaluations to reach the optimal solution. Hybrid approach can produce high quality clusters with small standard deviation on selected datasets compared to other methods.

The combination of the KM-GSA with NM-PSO is proposed. This Hybrid combination improves the quality of data clustering and provides improvement over individual algorithm.

### A. EXPERIMENTAL RESULTS

Iris datasets are used to validate our proposed algorithm. Each dataset has a different number of clusters, data objects and features . These datasets have been used by many authors to compare and evaluate the performance of clustering algorithms in the literature and are described as follows: Iris dataset (n= 150,d=4,k=3): This dataset contains three classes of 50 objects each, where each class refers to a type of iris flower. There are 150 random samples of iris flowers with four numeric attributes in this dataset. These attributes are sepal length and width in cm, petal length and width in cm. There are no missing values for attributes.

### B. PERFORMANCE MEASURE

On the iris data sets the intra cluster obtained for KM algorithm, KM-GSA algorithm and KM-GSA-NM-PSO algorithm respectively, which are much better than the results than the results obtained by the individual algorithms, which means than the K-means algorithm may trap in local optima in some cases while GSA-KM can converge to the optimal solution in most cases in comparison with original GSA alone. PSO has the convergence problem than all algorithms in general. This is overcome by Nelder-Mead algorithm.

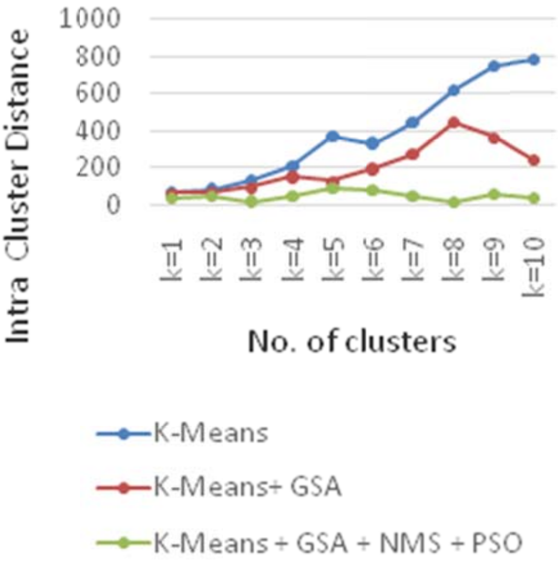

Figure 2.Comparison of InterCluster Distance

The Nelder Mead algorithm provides efficient local search procedure but its convergence is extremely sensitive to the chosen starting point. The percentage increase between the algorithms is as shown fig.4.2

TABLE 1: COMPARISON OF INTRACLUSTER DISTANCE

| K Value | K-Means | K-Means+ GSA | K-Means + GSA + NM + PSO |
|---|---|---|---|
| K=1 | 70.7177 | 67.0894 | 41.2553 |
| K=2 | 86.7219 | 70.746 | 53.8702 |
| K=3 | 131.6436 | 99.4696 | 22.4053 |
| K=4 | 206.6357 | 157.2087 | 51.0000 |
| K=5 | 366.3687 | 128.5151 | 94.3451 |
| K=6 | 331.027 | 200.8252 | 79.3788 |
| K=7 | 442.3062 | 279.5675 | 52.9244 |
| K=8 | 617.0131 | 449.479 | 17.3200 |
| K=9 | 746.2693 | 368.9369 | 57.6714 |
| K=10 | 776.1084 | 245.9955 | 36.0555 |

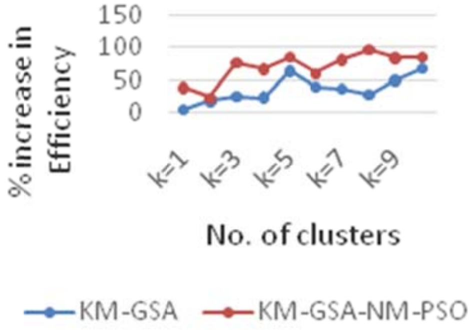

Figure3. Efficiency comparison

Table 2.Efficiency comparison

| No. of Clusters | KM vs. KM-GSA | KM-GSA vs. KM-GSA-NM-PSO |
|---|---|---|
| K=1 | 5 | 38 |
| K=2 | 18 | 23 |
| K=3 | 24 | 77 |
| K=4 | 23 | 67 |
| K=5 | 64 | 84 |
| K=6 | 39 | 60 |
| K=7 | 36 | 81 |
| K=8 | 27 | 96 |
| K=9 | 50 | 84 |
| K=10 | 68 | 85 |

The efficiency comparison table shows that the KM vs KM-GSA and KM-GSA vs. KM-GSA-NM-PSO in table 4.2. The efficiency is computed based on the intra cluster distance obtained on each individual algorithm.

The KM-GSA-NM-PSO algorithm has produced the highest quality solutions in terms of the best intra cluster distances on all the test datasets. Moreover, the standard deviation of solutions found by KM-GSA-NM-PSO is the smallest, which means that KM-GSA-NM-PSO can find a near optimal solution in most of the runs while other algorithms may trap local optima insome of the runs. In other words, the results confirm that the proposed algorithm is viable and robust. In terms of the number of function evaluations, the k-means algorithm needs the least number of evaluations compared to the individual algorithms. In other words, in the KM-GSA-NM-PSO algorithm, the GSA method starts from a good initial state due to the use of the output of k-means and consequently reaches the optimal solution faster than the pure GSA because in the pure GSA alone, all candidate solutions are generated randomly and the quality of the initial population is not as good as the initial population in the GSA-KM algorithm.

The KM-GSA algorithm tends to converge faster than the PSO as it requires fewer function evaluations, but it usually results in less accurate clustering. One can take advantage of its speed at the inception of the clustering process and leave accuracy to be achieved by other methods at a later stage of the process. This statement shall be verified in later sections of this paper by showing that the results of clustering by PSO and NM–PSO can further be improved by seeding the initial population with the outcome of the K-means algorithm More specifically, the hybrid algorithm first executes the K-means algorithm, which terminates when there is no change in centroid vectors. In the case of K– PSO, the result of the K-means algorithm is used as one of the particles, while the remaining 5N-1 particles are initialized randomly.. In the case of KM-GSA–NM–PSO, randomly generate 3N particles, or vertices as termed in the earlier introduction of NM, and NM–PSO is then carried out to its completion.

```
Step 1: k-means method
1.1. Randomly choose k centroids from dataset for desired clusters
1.2. Assign each data object to the cluster with the closest centroid
1.3. Update the centroids by calculating the mean values of objects within clusters
1.4. Repeat steps 1.2 and 1.3 until termination criteria are met
Step 2: Generate an initial population of size S ({P1, P2, …, Ps}).
2.1 P1=k-means (dataset) // Use output of k-means as one of the candidate solutions
2.2 P2=min (dataset) // Generate a candidate solution using the minimum of the dataset
2.3 P3=mean (dataset) // Generate a candidate solution using the mean of the dataset
2.4 P4=max (dataset) // Generate a candidate solution using the maximum of the dataset
2.5 P5…Ps=random (dataset) // Generate all other candidate solutions randomly
Step 3: GSA method
3.1. Calculate the fitness function for all of the particles (candidate solutions)
3.2. Calculate M, F and a for all of the particles based on Eq. (6, 8 and 9) as described in the GSA algorithm
3.3. Update the velocity and position of particles based on Eq.(10 and 11) as described in the GSA algorithm
3.4. If termination criteria are met (i.e., the predefined number of iteration is reached or the fitness function is satisfied) output the best particle, which has the best value for the fitness function as the final solution; otherwise return to step 3.1.
Step 4: Nelder-Mead method
4.1. Initialization : Generate a population of size 3N +1.
4.2. Evaluation & Ranking :Evaluate the fitness of each particle. Rank them on the basis of fitness.
4.3. Simplex Method :Apply NM operator to the top N +1 particles and replace the (N +1)th  particle with the update.
Step 5:  PSO Method
5.1.Apply PSO operator for updating the remaining 2N p articles.
5.2.Selection: From the population select the global best particle and the neighborhood best particles.
5.3Velocity Update: Apply velocity update to the 2N particles with worst fitness according equations (3) and (4).
5.4. If the termination conditions are not met, go back to 4.2.
```

Figure 4. Main steps involved in KM-GSA-NM-PSO algorithm

## V. CONCLUSION

A hybrid method (coded as GSA-KM) that is based on a gravitational search algorithm (GSA) and k-means algorithm is used in clustering data objects. It tries to exploit the merits of two algorithms simultaneously, where the k-means is used in generating the initial solution and the GSA is employed as an improvement algorithm. The performance of the existing algorithm is compared with other approaches. The comparisons how that the existing algorithm over comes the short comings of k-means and GSA alone. It requires minimum number of function evaluations to reach the optimal solution. Moreover, the proposed approach will be combination of the existing algorithm with NM-PSO which can produce high quality clusters with small standard deviation on selected datasets compared to other methods. In future research, the proposed method may be applied to other applications, such as image segmentation and university time tabling. The combination of the KM-GSA-NM-PSO with other heuristic approaches and their application to data clustering is another research direction.

## VI. REFERENCES

[1] Abdolreza Hatamlou, Salwani Abdullah, Hossein Nezamabadi-pour," A combined approach for clustering based on K-means and gravitational search algorithms " Swarm and Evolutionary Computation, Volume 6, October 2012, pp. 47-52,2012.

[2] Yi-Tung Kao, Erwie Zahara, I-Wei Kao, "A hybridized approach to data clustering, Expert Systems with Applications", Volume 34, Issue 3, pp. 1754-1762, 2008.

[3] L.E. Agustı´n-Blas, S. Salcedo-Sanz, S. Jiménez-Fernández, L. Carro-Calvo, J. Del Ser, J.A. Portilla-Figueras, "A new grouping genetic algorithm for clustering problems", Expert Systems with Applications, Volume 39, Issue 10, pp.9695-9703, 2012.

[4] Minghao Yin, Yanmei Hu, Fengqin Yang, Xiangtao Li, Wenxiang Gu, "A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering", Expert Systems with Applications, Volume 38, Issue 8, pp. 9319-9324, 2011.

[5] Hua Jiang, Shenghe Yi, Jing Li, Fengqin Yang, Xin Hu, "Ant clustering algorithm with K-harmonic means clustering", Expert Systems with Applications, Volume 37, Issue 12, pp. 8679-8684, 2010.

[6] Esmat Rashedi, Hossein Nezamabadi-pour, Saeid Saryazdi, "GSA: A Gravitational Search Algorithm", Information Sciences, Volume 179, Issue 13, pp. 2232-2248,2009.

[7] J. Kennedy, R. Eberhart," Particle swarm optimization", in: Neural Networks. Proceedings, IEEE International Conference on, 1995.

[8] P. Jin, Y.L. Zhu, K.Y. Hu, "A clustering algorithm for data mining based on swarm intelligence", in: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC, 2007.

[9] B. Saglam, et al., A mixed-integer programming approach to the clustering problem with an application in customer segmentation, European Journal of Operational Research 173 (3) (2006) 866–879.

[10] A.K. Jain, Data clustering: 50 years beyond Kmeans, Pattern Recognition Letters 31 (8) (2010) 651–666.

[11] C. Ching-Yi, Y. Fun, Particle swarm optimization algorithm and its application to clustering analysis. in Networking, Sensing and Control, 2004 IEEE International Conference on, 2004.

[12] E.W. Forgy, Cluster analysis of multivariate data: efficiency versus interpret-ability of classifications, Biometrics 21 (1965) 2.

[13] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, New York, 1990.

[14] Y.-T. Kao, E. Zahara, I.W. Kao, A hybridized approach to data clustering, Expert Systems with Applications 34 (3) (2008) 1754–1762.

[15] S.Z. Selim, M.A. Ismail, K-means-type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6 (1) (1984) 81–87.

[16] S.Z. Selim, K. Alsultan, A simulated annealing algorithm for the clustering problem, Pattern Recognition 24 (10) (1991) 1003–1008.

[17] K.S. Al-Sultan, A Tabu search approach to the clustering problem, Pattern Recognition 28 (9) (1995) 1443–1451.

[18] C.S. Sung, H.W. Jin, A tabu-search-based heuristic for clustering, Pattern Recognition 33 (5) (2000) 849–858.

[19] U. Maulik, S. Bandyopadhyay, Genetic algorithm-based clustering technique, Pattern Recognition 33 (9) (2000) 1455–1465.

[20] K. Krishna, M. Narasimha Murty, Genetic K -means algorithm, IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics) 29 (3) (1999)433–439.

[21] A.K. Qin, P.N. Suganthan, Kernel neural gas algorithms with application to cluster analysis, in: Proceedings—International Conference on Pattern Recog- nition, 2004.

[22] A.K. Qin, P.N. Suganthan, A robust neural gas algorithm for clustering analysis, in: Proceedings of International Conference on Intelligent Sensing and Information Processing, ICISIP 2004, 2004.

[23] P.S. Shelokar, V.K. Jayaraman, B.D. Kulkarni, An ant colony approach for clustering, Analytica Chimica Acta 509 (2) (2004) 187–195.

[24] D. Karaboga, C. Ozturk, A novel clustering approach: artificial bee colony (ABC) algorithm, Applied Soft Computing 11 (1) (2011) 652–657.

[25] M. Fathian, B. Amiri, A. Maroosi, Application of honey-bee mating optimiza- tion algorithm on clustering, Applied Mathematics and Computation 190 (2) (2007) 1502–1513.

[26] A. Hatamlou, S. Abdullah, M. Hatamlou, Data clustering using big bang-big crunch algorithm, in: Communications in Computer and Information Science, 2011, pp. 383–388.