International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

K-harmonic means Data Clustering using Combination of Particle Swarm Optimization and Tabu Search

Tahereh Aghdasi[1*], Javad Vahidi[2], Homayoon Motameni[3] and Mohammad Madadpour Inallou[4]

[1]Islamic Azad University, Science and Research Branch of Ayatollah Amoli, Amol, Iran

[2]Iran University of Science and Technology, Tehran, Iran

[3]Islamic Azad University, Sari, Iran

[4]Young Researchers and Elites Club, Islamic Azad University, West Tehran Branch, Tehran, Iran

[*]Corresponding Author's Email: *tahere.aghdasi@gmail.com*

Abstract

Clustering is one of the widely used techniques for data analysis. Also it is a tool to discover structures from inside of data without any previous knowledge. K-harmonic means (KHM) is a center-based clustering algorithm which solves sensitivity to initialization of the centers which is the main drawback of K-means (KM) algorithm, but, both KM and KHM converge to local optimal. In this paper, a hybrid data clustering algorithm based on KHM is proposed called PSOTSKHM, using Particle Swarm Optimization (PSO) algorithm as a stochastic global optimization technique and Tabu Search (TS) algorithm as a local search method. This algorithm makes full use of the advantages of three algorithms. The proposed algorithm has been compared with KHM, PSOKHM and IGSAKHM algorithms on four real datasets and the obtained results show the superiority of suggested algorithm in most cases.

Keywords: Data clustering, K-harmonic means, Particle Swarm Optimization, Tabu Search.

1. Introduction

Clustering is a method of unsupervised learning. It organizes data instances into disjointed clusters such that the objects in the same cluster are similar to each other and objects in different clusters are as dissimilar as possible [3]. It has been widely used in

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

various fields and many applications such as machine learning, data mining, pattern recognition, image analysis and bioinformatics [14, 19]. The existing clustering algorithms can be basically classified into two categories: hierarchical clustering and partitional clustering. Hierarchical clustering approach aims at grouping data through repeated cluster separation or agglomeration. Partitional clustering approach attempts to directly decompose data into disjoint clusters based on an objective function such as minimizing the distance between data points and cluster centers [5].

The classic K-means algorithm (KM) is the most popular clustering algorithm due to it's simplicity and efficiency but it is very sensitive to initialization, and is easily trapped in local optimal. Recently much work was done to overcome these problems [7, 9]. As an improved version of KM, the K-harmonic means (KHM) algorithm was first proposed by Zhang et al [4] and modified by Hammerly and Elkan [8]. Similar to KM, KHM partitions data objects into k clusters while KHM uses the harmonic average of the distances from each data point to all cluster centers to find the clusters. Therefore, KHM is insensitive to the initialization of the centers, however it runs into local optimal [14].

In more references, for solving this problem, some heuristic clustering algorithms have been used. A method is suggested based on genetic algorithm called Genetic K-means algorithm (GKA) to solve the problem of clustering which defines a basic mutation operator to clustering. Using finite Markov chain theory, it has been proved that GKA converge to the global optimum. It is also observed that GKA searches faster than some of the other evolutionary algorithms used for clustering [12].

Shelokar et al an algorithm is provided based on ant colony algorithm. Results of this algorithm compared with other metaheuristics algorithms such as GA, TS and SA which showed that the performance of the proposed algorithm was better than other algorithms [16]. Hua et al was proposed a new algorithm using the Ant clustering algorithm with K-harmonic means clustering called ACAKHM [9]. A new hybrid data clustering algorithm based on KHM And IGSA, called IGSAKHM, was proposed by Minghao et al [14]. Gangor and Anler proposed a hybrid method with SA algorithm to solve the problem of

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

convergence to local optimal of KHM [22], also they suggested other hybrid algorithm using KHM and TS method called TABUKHM [23].

A new algorithm based on PSO and KHM was proposed by yang et al to optimize KHM algorithm called PSOKHM. This algorithm helps the KHM clustering escape from local optimal and overcomes the shortcoming of the slow convergence speed of the PSO algorithm [7]. Zhang et al proposed a Tabu Search Particle Swarm Optimization (TSPSO) algorithm. They proved that the TSPSO is superior to GA, TS, and PSO [20]. Also they will apply the TSPSO algorithm to cluster analysis problem [19].

In this paper a hybrid clustering algorithm based on KHM is suggested using the PSO and TS algorithms which uses the global search ability of PSO and local search ability of TS and enables to countervail the weaknesses of each three algorithms and find better performance in less time. Four real-life data sets and three algorithms were used for experiment and the results illustrate the superiority of the proposed clustering algorithm in most cases.

The rest of article is classified followed as: overall description of clustering problem is presented in section 2. A description of KHM algorithm is given in section 3. In section 4 and 5 PSO and TS algorithms are briefly reviewed and our proposed algorithm is introduced in section 6, then experimental studies are presented in section 7 and in section 8 conclusions are provided and look at ideas about the proposed algorithm.

2. Clustering problem

As it is pointed, clustering is organizing similar objects in the same cluster based on a similarity metric between two objects. The total clustering process include following steps [2]:

1. Pattern preparation which include feature extraction and/or selection,
2. Definition of a similarity measure appropriate to the data domain,
3. Clustering or grouping process,
4. Data abstraction and cluster analysis (if needed),

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

5. Validation of output.

Because of the variety of feature types and scales, The similarity measure must be chosen carefully. The most used distance metric is Euclidean distance. It is a special case (p=2) of the Minkowski metric which is defined as [2]:

$$d_p(x_i, x_j) = \sqrt[p]{\sum_{k=1}^{d} \left| x_{i,k} - x_{j,k} \right|^p} = \left\| x_i - x_j \right\|_p \tag{1}$$

Where $d_p(x_i, x_j)$ shows dissimilarity measure between object i and object j, and $x_{i,k}$ denotes the value of kth attribute of object i Thus, the number of an object's attributes are denoted by d.

3. K-Harmonic means algorithm

As noted previously, dependency of KM clustering results to the clusters initialization is a primary problem of this algorithm. KHM clustering can solve this problem by replacing the minimum distance from a data point to the centers, used in KM, by the harmonic mean of the distance from each data point to all centers [14]. Also in contrast to the KM algorithm which gives equal weight to all of the data points, KHM gives dynamic weight to each data point every time and uses from membership function. The following notations (modes) are used for KHM algorithm:

X =(x1, ... ,xn): the data should be clustered.

C =(c1, ... , ck): the cluster centers set.

m(cj|xi): the membership function which define the proportion of data point that belongs to cj center.

w(xi): the weight function which define the amount of xi effects in re-computing of the center parameters in the next iteration.

Steps of basic algorithm for KHM is following as [9, 14]:

1. Initialize the algorithm by random selection of the centers.

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

2. Calculation of objective function value which is following to:

$$KHM(X,C)=\sum_{i=1}^{n}\frac{K}{\sum_{j=1}^{K}\frac{1}{\|x_i-c_j\|^p}} \tag{2}$$

Where harmonic means is used to measuring distances and p is an input parameter (p>=2).

3. For each data point x_i, membership function is calculated in each center c according to:

$$m(c_j \mid x_i)=\frac{\|x_i-c_j\|^{-p-2}}{\sum_{j=1}^{K}\|x_i-c_j\|^{-p-2}} \tag{3}$$

4. For each data point x_i it's weight $w(x_i)$ calculated as:

$$w(x_i)=\frac{\sum_{j=1}^{K}\|x_i-c_j\|^{-p-2}}{\left(\sum_{j=1}^{K}\|x_i-c_j\|^{-p}\right)^2} \tag{4}$$

5. For each center c_j, re-compute it's location from all data points x_i according to their memberships and weights:

$$c_j=\frac{\sum_{i=1}^{n}m(c_j \mid x_i)\,w(x_i)\,x_i}{\sum_{i=1}^{n}m(c_j \mid x_i)\,w(x_i)} \tag{5}$$

6. Repeat steps 2 to 5 until it reaches the predefined number of iterations or $KHM(X,C)$ does not change significantly.
7. Allocate the point xi to the cluster j with the biggest $m(c_j \mid x_i)$.

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

4. Particle Swarm Optimization

PSO is a population-based stochastic optimization technique developed by Eberhart and Kennedy [17]. PSO was inspired by social behavior of bird flocking or fish schooling. each single solution is considered as a particle and each particle has fitness value, position and velocity that attempts to move toward a better solution [15]. PSO algorithm procedure is described as follows [5]:

1. Particle initialization: An initial swarm of particles is generated in search space. Usually, the population size is decided by the dimension of problems.
2. Velocity and position update: In every iteration, a new velocity value for each particle is calculated based on it's current velocity, the distance from it's previous best position, and the distance from the global best position by equation (6). Then the new position of the particle is calculated using the new velocity value by equation (7).

$$V_{id}^{new} = \omega \times V_{id}^{old} + c_1 \times rand_1 (\mathrm{P}_{id} - x_{id}^{old}) + c_2 \times rand_2 (\mathrm{P}_{gd} - x_{id}^{old}) \tag{6}$$

$$x_{id}^{new} = x_{id}^{old} + V_{id}^{new} \tag{7}$$

V is particle's velocity, x is particle's position, c_1 and c_2 are acceleration coefficients that are conventionally set to a fixed value between 0 and 2. P_{id} is the previous individual best position of a particle and P_{gd} is the current global best position. ω is an inertia weight that larger inertia ensures a more effective global search of particles and smaller inertia weight means a more efficient local search [11]. According to this theory, Shi and Eberhart proposed a strategy which inertia weight linearly decreases with the increase number of iterations Expressed as follows [21]:

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

$$\omega = \omega_{\max} - k \times \left(\omega_{\max} - \omega_{\min}\right) / gen \tag{8}$$

Which $\omega_{\max}$ calls as the maximum inertia weight, $\omega_{\min}$ calls the minimum inertia, gen calls the total number of iterations for the algorithm, k calls the current number of iterations for the algorithm.

3. Evaluation and update of best locations: The fitness value of each particle is calculated by the objective function. The values of P_{id} and P_{gd} are then evaluated and replaced if better particle best position or global best position is obtained.
4. Termination: Repeat steps 2 and 3 until the termination condition is met.

5. Tabu Search

Tabu Search is a search method for combinatorial optimization problems which is proposed by Glover in 1989 [6]. TS uses a local or neighborhood search procedure to iteratively move from a solution S to a solution S1 in the neighborhood of S until some stopping criterion has been satisfied. Perhaps the most important aspect of TS is it's memory structure which is called ''Tabu List''. In it's simplest form, a tabu list contains the solutions that have been visited less than n moves ago, where n is the length of tabu list.

The basic components of the TS algorithm as follows [13]:

1. Configuration is an assignment of values to variables. It is a solution to the optimization problem.
2. Move is a specific procedure for getting a trial solution Which is feasible to the optimization problem that is related to the current configuration.
3. Neighborhood is the set of all neighbors, which are the "adjacent solutions" that can be reached from any current configuration. It may also include neighbors that do not satisfy the given customary feasible conditions.

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

4. Candidate subset is a subset of the neighborhood. It is to be examined instead of the entire neighborhood, especially for large problems where the neighborhood have many elements.
5. Tabu restrictions are constraints that prevent the chosen Moves to be reversed or repeated. They play a memory Role for the search by making the forbidden moves as tabu. The tabu moves are stored in a list, called tabu list.
6. Aspiration criteria are rules that determine when the tabu restrictions can be overridden, thus removing a tabu classification otherwise applied to a move. If a certain move is forbidden by some tabu restrictions then the aspiration criteria, when satisfied, can make this move allowable.

6. The proposed algorithm

The KHM algorithm converges faster than the PSO algorithm but usually it trapped in local optimum. Although PSO could be effective to solving this problem but it converges too slow or even converges to local optimal [18].

In this research we develop an algorithm by combining PSO and TS called PSOTSKHM to speed up the convergence of PSO and finding the good result in a minimum time. In PSO function of our algorithm, firstly amount of particle's step length is considered to be high to increase ability of global search. After executing PSO function we select the best particle then we apply TS function to search it's neighbor carefully to obtaining the best results. Considering that our algorithm conducts both global search and local search in each of iterations, the probability of finding the better solution increases.

A particle is a vector of real numbers with $k \times d$ dimension, so that k is the clusters number and d is the dimension of data to be clustered. In Figure 1 a model of particle is represented.

x_{11}	x_{12}	...	x_{1d}	...	x_{k1}	x_{k2}	...	x_{kd}

Figure 1: The representation of a particle

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

Two following swap strategy are used to creating neighborhood in TS function:

1. A center of cluster is selected randomly representative c_j, then we choose the closest pattern x_j to the c_j and assign the x_j as a new cluster replace to c_j. Remaining centers are the same.
2. A center of cluster is selected randomly representative c_j, then we choose the pattern x_j randomly and assign the x_j as a new cluster replace to c_j. Remaining centers are the same.

Steps of suggested algorithm are following as:

1. Set the initial parameters
2. Initialize the population in the number of p_{size} randomly
3. Calculating the fitness function of each particle using KHM function
4. Executing PSO and update the position and velocity of each particle in a considered number.
5. Executing TS for the best particle of PSO
6. Updating the "p_{best}" and "g_{best}" particles
7. Apply KHM for the best particle in the result.
8. Repeat Step 3 to Step 7 until the final criterion was met.
9. Output representation of obtained results.

6.1. Parameter tuning

According to section 4 about ω parameter In PSO algorithm, with setting ω, a balance could be created between local and global search. Larger inertia affords a more effective global search of particles [11]. In this article, it is tried to increase the ability of global searching of PSO. After of three iterations for PSO we run TS with one iteration then use KHM with four iterations. The PSO parameters set in our algorithm show in table 1:

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

Table 1: The parameters set in PSO function

Parameter	Value
p_{size}	18
ω	0.9
ω_{max}	0.6
ω_{min}	0.4
c_1	1.496
c_2	1.496

The abbreviations of parameters which are used in the TS function are following as:

MTLS: Maximum size of tabu list

NTS: Number of trial solutions

P: Probability threshold

TLL: Tabu list length

The TS parameters set in our algorithm show in table 2:

Table 2: The parameters set in TS function

Parameter	Value
p	0.95
MTLS	10
NTS	20

7. Experimental studies

We test our proposed algorithm on four real data sets and compared with three algorithms. These data sets include Iris, Glass, Breast-Cancer-Wisconsin (denoted as Cancer), and Contraceptive Method Choice (denoted as CMC), which in [7] and [14] have been applied. Table 3 lists the details of the data sets. The experimental results are averages of 10 runs. The algorithms are implemented using matlab software and ran on a CPU 2. 3 GHz with 4.00 GB RAM memory.

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

Table 3: Characteristics of data sets considered.

Name of data set	No. of classes	No. of features	Size of data set (size of classes)
Iris	3	4	150 (50, 50, 50)
Cancer	2	9	683 (444, 239)
CMC	3	9	1473 (629, 334, 510)
Glass	6	9	214 (70, 17, 76, 13, 9, 29)

7.1. Data sets

1. Fisher's iris data set (n =150, d =4, k =3), which consists of three differents pecies of iris flower: Iris Setosa, IrisVersicolour and IrisVirginica. For each species, 50 samples with four features (sepal length, sepal width, petal length, and petal width) were collected.
2. Glass (n =214, d =9, k =6), which consists of six different types of glass: building windows float processed (70 objects), building windows non-float processed (76 objects), vehicle windows float processed (17 objects), containers (13 objects), tableware (9 objects), and headlamps (29objects). Each type has nine features, which are refractive index, sodium, magnesium, aluminum, silicon, potassium, calcium, barium, and iron.

3. Wisconsin breast cancer (n =683, d =9, k =2), which consists of 683 objects characterized by nine features: clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. There are two categories in the data: malignant (444 objects) and benign (239 objects).

4. Contraceptive Method Choice (n =1473, d =9, k =3): This data set is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

married women who either were not pregnant or did not know if they were at the time of interview. The problem is to predict the choice of current contraceptive method (no use has 629 objects, long-term methods have 334 objects, and short-term methods have 510 objects) of a woman based on her demographic and socioeconomic characteristics.

7.2. Experimental results

In this part, we compare the performances of KHM, PSOKHM, IGSAKHM methods with proposed algorithms. The following two criteria have used to measuring The quality of algorithms [7, 14].

1. The value of KHM function, as defined in equation (2). Clearly, the smaller the sum is, the higher the quality of clustering is.

2. The F-Measure value, it is a quality measure which uses the ideas of precision and recall from information retrieval [1, 10]. Each class of i as shown by the class labels in the data set is considered as the desired set of n_i items for a query, each cluster j which produced by the algorithm is considered as the set of n_j items retrieved for a query, n_{ij} is the numbers of class i members inside cluster j. For each class i and cluster j precision and recall are defined as equation (9) and (10) and the value of corresponding F-Measure is given by equation (11) where set b=1 to have equal weighting for $p(i,j)$ and $r(i,j)$. Equation (12) shows the value of total F-Measure for the data set with the size of n. It is cleare that, the bigger value of F-Measure is, the higher the quality of clustering is.

$$p(i,j) = \frac{n_{ij}}{n_j} \tag{9}$$

$$r(i,j) = \frac{n_{ij}}{n_i} \tag{10}$$

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

$$F(i,\mathrm{j}) = \frac{(\mathrm{b}^2+1).p(i,j).r(i,j)}{\mathrm{b}^2.p(i,j)+r(i,j)} \tag{11}$$

$$F = \sum_i \frac{n_i}{n} \max_j \{F(i,j)\} \tag{12}$$

Results of experiments on the KHM algorithm show that p is a key parameter to reaching to good objective function values. For this reason in this research our experiments conducted with various p values. Tables 4 - 6 are results of implementation of algorithms for different value of 2.5, 3, 3.5 and the average of KHM fitness functions, F-Measures and runtimes in 10 independent runs of each algorithm have been compared. Also the runtimes of the algorithms are shown in tables and standard deviations is represented in brackets. The best results are shown bold and the second best results are shown italic.

Table 4: Results of KHM, PSOKHM, IGSAKHM, and the proposed algorithm (PSOTSKHM) when p=2.5

	KHM	PSOKHM	IGSAKHM	PSOTSKHM
Iris				
KHM	149.333(0.000)	*149.058(0.074)*	*149.058(0.000)*	**148.890(0.011)**
F-Measure	0.750(0.000)	0.753(0.005)	*0.763(0.000)*	**0.8859(0.000)**
Runtime	**0.192(0.008)**	1.842(0.005)	1.577(0.002)	*1.373(0.210)*
Cancer				
KHM	60189(0)	*59844(22)*	*59844(0)*	**57010(35)**
F-Measure	*0.829(0.000)*	*0.829(0.000)*	*0.829(0.000)*	**0.9617(0.000)**
Runtime	**2.017(0.009)**	9.525(0.013)	7.509(0.007)	*5.878(0.955)*
CMC				
KHM	96520(0)	*96193(25)*	*96193(52)*	**96174(10.17)**
F-Measure	0.335(0.000)	0.333(0.002)	**0.488(0.000)**	*0.462(0.000)*
Runtime	**8.639(0.009)**	39.825(0.072)	31.563(0.012)	*14.317(2.291)*
Glass				
KHM	1203.554(16.231)	1196.789(0.439)	**1180.756(0.134)**	*1193.328(0.262)*
F-Measure	0.421(0.011)	0.424(0.003)	*0.454(0.000)*	**0.703(0.002)**
Runtime	*4.064(0.010)*	17.669(0.018)	15.910(0.010)	**3.489(0.552)**

International Journal of Mechatronics, Electrical and Computer Technology

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

Table 5: Results of KHM, PSOKHM, IGSAKHM, and the proposed algorithm (PSOTSKHM) when p=3

	KHM	PSOKHM	IGSAKHM	PSOTSKHM
Iris				
KHM	126.517(0.000)	*125.951(0.052)*	*125.951(0.000)*	**125.940(0.043)**
F-measure	0.744(0.000)	0.744(0.000)	*0.751(0.000)*	**0.891(0.000)**
Runtime	**0.190(0.007)**	1.826(0.009)	1.650(0.004)	*1.349(0.206)*
Cancer				
KHM	119458(0)	*117418(237)*	*117418(55)*	**112562(275)**
F-measure	0.834(0.000)	0.834(0.000)	*0.847(0.000)*	**0.964(0.000)**
Runtime	**2.027(0.007)**	9.594(0.023)	7.91(0.002)	*5.952(0.970)*
CMC				
KHM	187525(0)	**186722(111)**	**186722(94)**	*186840(54)*
F-measure	0.303(0.000)	0.303(0.000)	**0.472(0.000)**	*0.4617(0.000)*
Runtime	**8.627(0.009)**	39.485(0.056)	32.107(0.034)	*15.164(2.502)*
Glass				
KHM	1535.198(0.000)	1442.847(35.871)	*1400.950(0.630)*	**1394.753(0.697)**
F-measure	0.422(0.000)	0.427(0.003)	*0.442(0.000)*	**0.713(0.000)**
Runtime	*4.042(0.007)*	17.609(0.015)	15.958(0.001)	**3.982(0.630)**

Table 6: Results of KHM, PSOKHM, IGSAKHM, and the proposed algorithm (PSOTSKHM) when p=3.5

	KHM	PSOKHM	IGSAKHM	PSOTSKHM
Iris				
KHM	113.413(0.085)	*110.004(0.260)*	*110.004(0.002)*	**109.819(0.175)**
F-measure	*0.770(0.024)*	0.762(0.004)	0.766(0.000)	**0.891(0.000)**
Runtime	**0.194(0.008)**	1.873(0.005)	1.587(0.004)	*1.310(0.198)*
Cancer				
KHM	243440(0)	*235441(696)*	236125(15)	**228387(906)**
F-measure	0.832(0.000)	0.835(0.003)	*0.862(0.000)*	**0.966(0.000)**
Runtime	**2.072(0.008)**	9.859(0.015)	31.521(0.009)	*5.596(0.914)*
CMC				
KHM	381444(0)	**379678(247)**	380183(16)	*380048.191(220)*
F-measure	0.332(0.000)	0.332(0.000)	**0.506(0.000)**	*0.459(0.000)*
Runtime	**8.528(0.012)**	42.701(0.250)	31.521(0.009)	*16.632(2.645)*
Glass				
KHM	1871.812(0.000)	*1857.152(4.937)*	*1857.152(0.035)*	**1843.133(4.107)**
F-measure	0.396(0.000)	0.396(0.000)	*0.420(0.000)*	**0.663(0.000)**
Runtime	*4.056(0.008)*	17.651(0.013)	15.799(0.003)	**3.502(0.555)**

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

Conclusions

In this article, an algorithm of data clustering was developed based on KHM algorithm using combination of PSO and TS, called PSOTSKHM. Four well-known standard data sets, three algorithms namely KHM, IGSAKHM and PSOKHM and two quality measurement criteria were used to evaluate the algorithm. Experimental results show that the proposed algorithm for reaching to the best solution, has more convergence speed than other algorithms and doesn't trapped in local optimal. The improvement of two evaluation criteria, in most cases, represents the more efficiency of proposed algorithm than three other algorithms. Also the time of running compared with IGSAKHN and PSOKHM has been improved. To completing done work, various researches could be proposed such as Apply a method for determining the number of clusters automatically, use from other metaheuristics algorithms to combining with KHM and apply the PSOTSKHM to use in applications such as image processing, data mining and crime prediction.

References

[1] A. Dalli, "Adaptation of the F-measure to cluster based lexicon quality evaluation", in Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?, (2003), pp. 51-56.

[2] A.K. Jain, M.N. Murty and P.J. Flynn, "Data clustering: a review", ACM Computing Surveys (CSUR), Vol. 31, 3 (1999), pp. 264-323.

[3] B. Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer, (2007).

[4] B. Zhang, M. Hsu and U. Dayal, "K-harmonic means-a data clustering algorithm", Hewlett-Packard Labs Technical Report HPL-1999-124, (1999).

[5] C.-Y. Tsai and I.-W. Kao, "Particle Swarm Optimization with selective particle regeneration for data clustering", Expert Systems with Applications, Vol. 38, 6 (2011), pp. 6565-6576.

[6] F. Glover and M. Laguna, "Tabu Search", Springer, (1997), Vol. 22.

[7] F. Yang, T. Sun, and C. Zhang, "An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization", Expert Systems with Applications, Vol. 36, 6 (2009), pp. 9847–9852.

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

[8] G. Hamerly and C. Elkan. "Alternatives to the k-means algorithm that find better clusterings", In Proceedings of the Eleventh International Conference on Information and knowledge Management, (2002), pp. 600-607, ACM.

[9] H. Jiang, S. Yi, J. Li, F. Yang, and X. Hu, "Ant clustering algorithm with K-harmonic means clustering", Expert Systems with Applications, Vol. 37, 12 (2010), pp. 8679–8684.

[10] J. Makhoul, F.Kubala, R. Schwartz and R. Weischedel, "Performance Measures for Information Extraction", Proceedings of DARPA Broadcast News Workshop, (1999), pp. 249-252.

[11] K.-h. WANG, H. NIU, Y.-n. ZHANG and X.-p. ZHANG, "Particle Swarm Optimization with adaptive parameters and boundary constraints", Electronic Design Engineering, Vol. 19, 21 (2011), pp. 46-52.

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

[13] M.K. Ng and J.C. Wong, "Clustering categorical data sets using Tabu search techniques", Pattern Recognition, Vol. 35, 12 (2002), pp. 2783-2790.

[14] M. Yin, Y. Hu, F. Yang, X. Li and W. Gu, "A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering", Expert Systems with Applications, Vol. 38, 8 (2011), pp. 9319–9324.

[15] N. Talbi and K. Belarbi, "Fast hybrid PSO and Tabu search approach for optimization of a fuzzy controller", IJCSI, International Journal of Computer Science, Vol. 8, 2 (2011), pp. 215-219.

[16] P. Shelokar, V.K. Jayaraman and B.D. Kulkarni, "An ant colony approach for clustering", Analytica Chimica Acta, Vol. 509, 2 (2004), pp. 187-195.

[17] R. Eberhart and J. Kennedy. "A new optimizer using particle swarm theory", in Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on, (1995), pp. 39-43, IEEE.

[18] V. Kalivarapu, J.-L. Foo and E. Winer, "Synchronous parallelization of Particle Swarm Optimization with digital pheromones", Advances in Engineering Software, Vol. 40, 10 (2009), pp. 975-985.

[19] Y. Zhang and L. Wu, "Tabu Search Particle Swarm Optimization used in cluster analysis", Journal of Science, Vol. 1, 1 (2012), pp. 6-12.

[20] Y. Zhang and L. Wu, "A hybrid TS-PSO optimization algorithm", Journal of Convergence Information Technology, Vol. 6, 5 (2011), pp. 169-174.

[21] Y. Shi and R. Eberhart. "A modified particle swarm optimizer", in Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on, (1998), pp. 69-73.

Vol. 4(11), Apr, 2014, pp 485-501, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org

[22] Z. Gungor and A. Unler, "K-harmonic means data clustering with simulated annealing heuristic", Applied Mathematics and Computation, Vol. 184, 2 (2007), pp. 199–209.

[23] Z. Gungor and A. Unler, "K-Harmonic means data clustering with tabu-search method", Applied Mathematical Modelling, Vol. 32, 6 (2008), pp. 1115–1125.