A NOVEL HYBRID DATA CLUSTERING ALGORITHM USING K-HARMONIC MEANS AND PARTICLE SWARM OPTIMIZATION WITH LEVY FLIGHT

R.Jensi

Dr.Sivanthi Aditanar College of Engineering, Tiruchendur
Tamilnadu, India

ABSTRACT

Clustering is the process of partitioning a set of data objects into groups/clusters so that the objects within a cluster are more similar to each other while objects in different clusters are dissimilar. Clustering is the popular data mining technique, which is widely used in several research areas. K-Means (KM) algorithm is the one of the widely used partitional clustering algorithm because of its simplicity and ease of implementation. However, KM algorithm faces the problem in initializing the cluster centers and hence caught in local optima. K-Harmonic Means (KHM) is more insensitive to cluster center initialization than KM algorithm. Particle Swarm Optimization (PSO) is a stochastic global optimization technique which avoids the trapping in local optima. In this paper, a novel hybrid data clustering method based on PSO with Levy Flight and KHM (PSOLF-KHM) is proposed. Levy Flight is a random walk, in which steps takes place on a discrete grid rather than on a continuous space, which provides high exploration in the search space. The PSOLF- KHM merges the benefits of KHM and PSOLF, which avoids the trapping from local optima and also discounts the slow convergence speed of PSO algorithm. The superiority of the proposed algorithm is attained when comparing the results with KHM and PSOKHM algorithms.

KEYWORDS

Data Clustering, K-Harmonic Means, Particle Swarm Optimization, Levy Flight

1. INTRODUCTION

Clustering is an unsupervised learning technique that groups the data items into a set of disjoint clusters so that the data objects in each cluster are more similar to each other while the objects in different clusters are dissimilar. Clustering algorithm is mainly classified into two categories: hierarchical and partitional clustering [1-2]. A nested set of cluster is organized into a tree by the method called Hierarchical clustering. The lowest level in the tree represents a set of clusters in which each item is in its unique cluster and at the highest level, all items belong to the same cluster. The hierarchical algorithms consist of two divisions namely agglomerative or divisive. Agglomerative algorithms, also referred as the bottom-up algorithms, are the ones that treat each object as a single cluster in the beginning and successively merge the pair of groups that are close to one another until all of the groups are grouped into one. Divisive algorithms, also referred as the top-down algorithms, which continues with the process in the same cluster where, in each upcoming iteration, a cluster will be split up by an algorithm called flat clustering algorithm recursively until each and every object is in a single cluster. On the other hand partitional clustering creates a set of non-overlapping clusters such that each and every data object should be available in exactly one cluster. All the above said methods require the desired number of clusters to create final set of clusters. The commonly used partitioning method is k-means.

The classic k-means algorithm is the widely used clustering method due to its simplicity and ease of implementation. However, k-means suffers from initialization of cluster centers and also trapped in local optima. As Zhang, Hsu, & Dayal, 1999 [3-4], K-Harmonic means is not sensitive to initial cluster center assignment, but it also traps in local optima. K-Harmonic means is integrated with nature inspired swarm intelligence algorithms [5].

In past decades, many nature-inspired evolutionary algorithms have been developed for solving most engineering design optimization problems, which are highly nonlinear, involving many design variables and complex constraints. These metaheuristic algorithms are attracted very much because of the global search capability and take less time to solve real world problems. Nature-inspired algorithms imitate the behaviors of the living things in the nature, so they are also called as Swarm Intelligence (SI) algorithms. These metaheuristic algorithms are used for data clustering [17].

Particle Swarm Optimization [6] was inspired by the social behavior of bird flocks or fish schooling. KHM and PSO were combined to perform data clustering [11]. PSO-KHM algorithm solved the problem of trapping in local optima and slow convergence speed behavior.

Kao, Zahara, & Kao (2008) [10] proposed a hybrid technique based on K-means, Nelder-Mead simplex search, and Particle Swarm Optimization (K-NM-PSO).

Ant Colony Optimization (ACO) was initially proposed by Marco Dorigo based on the behavior of ants aims at the search for an optimal pathway in a graph. A new hybrid algorithm based on ACO and KHM was proposed [12].

Minghao, Yanmei , Yang, Li & Wenxiang, (2011) [13] proposed a new hybrid clustering algorithm based on KHM and Gravitational Search Algorithm (GSA). The authors proved that the KHM-GSA algorithm was superior to KHM and PSOKHM algorithms.

Moreover, K-Harmonic Means algorithm was combined with simulated annealing [16] and Tabu Search algorithm [15] to find optimal cluster centers.

In this paper, a new hybrid algorithm using Particle Swarm Optimization with Levy Flight and K-Harmonic Means (PSOLF-KHM) is proposed. The remaining sections of this paper are organized as follows. Section 2 presents K-Harmonic means data clustering algorithm. Particle Swarm Optimization algorithm and Levy Flight is presented in Section 3 and Section 4 respectively. Then in Section 5 proposed algorithm PSOLF-KHM is explained. Section 6 discusses datasets and experimental results and Section 7 concludes the paper with fewer discussions.

3. K-HARMONIC MEANS CLUSTERING

Likewise K-means (KM), K-Harmonic Means is also a partitional and center based clustering algorithm and it was proposed by Zhang et.al. (1999) [3-4]. The difference between KM and KHM is that the KHM uses the sum over all data points of the harmonic average of the squared distance from a data point to all the centers. The KHM algorithm is insensitive to the initialization of cluster centers; however, it also leads to local optima [16].

The common notations used in the procedure of clustering [11,16] are given as:

$X=\{x_1,x_2\ldots,x_n\}$: the set of n data items to be clustered.
$C=\{c_1,c_2,\ldots c_k\}$: the set k cluster centers.
KHM(X,C): The objective function of the KHM algorithm.
m(cj/xi): The degree of membership of the point xi belongs to cluster j.
w(xi): The degree of influence value of the point xi to the position of center cj in the next iteration.

The KHM clustering algorithm is illustrated as follows:

1. Randomly initialize the cluster centers C.

2. Calculate objective function value according to

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

   where the input parameter is p and typically p≥2

3. Compute membership $m(c_j|x_i)$ for each data point x_i in each center c_j according to

$$m(c_j|x_i)=\frac{\|x_i-c_j\|^{-p-2}}{\sum_{j=1}^{k}\|x_i-c_j\|^{-p-2}} \qquad (2)$$

4. Compute weight $w(x_i)$ for each data point x_i according to

$$w(x_i)=\frac{\sum_{j=1}^{k}\|x_i-c_j\|^{-p-2}}{\left(\sum_{j=1}^{k}\|x_i-c_j\|^{-p}\right)^{2}} \qquad (3)$$

5. Compute new center location using the membership and weight of each data point:

$$c_j=\frac{\sum_{i=1}^{n}m(c_j|x_i)w(x_i)x_i}{\sum_{i=1}^{n}m(c_j|x_i)w(x_i)} \qquad (4)$$

6. Repeat steps 2-5 until maximum number of iterations is reached or until the value of KHM(X, C) does not vary significantly.
7.
8. Assign data point x_i to cluster j with the biggest $m(c_j|x_i)$.

4. Particle Swarm Optimization (PSO)

Particle Swarm Optimization **(**PSO) [5-6] is a population based stochastic optimization technique. It is inspired by the flocks of birds' behavior and also by the behavior of fish schooling. A PSO algorithm is started with a population of candidate solutions. This population is called a swarm and each and every solution is named a particle. At first these particles are assigned with a random initial position and also assigned with an initial velocity. A value obtained from the objective function is the position of each particle. These particles are moved around in the search-space. While moving in the search space, particles remember its own best position and also the position of the best solution in the swarm. . They are called pbest and gbest. The value pbest is the best solution that has achieved so far. The value gbest is the value that can be obtained so far by any particle in the population. After calculating the two best values, the particle updates its velocity and positions. When improved positions are being discovered these will then come to guide the movements of the swarm. In each iteration, particles are updated according to the following equations:

$$v_i(t+1)=\omega*(v_i(t)+l_1*rand*\left(pbest(t)-x_i(t)\right)+l_2*rand*(gbest-x_i(t))) \quad (5)$$

$$x_i(t+1)=x_i(t)+v_i(t+1) \quad (6)$$

where $v_i(t+1)$ is the velocity of ith particle at iteration t+1

$v_i(t)$ is the velocity of i^{th} particle at iteration t

$x_i(t+1)$ is the position of i^{th} particle at iteration t+1

$x_i(t)$ is the position of i^{th} particle at iteration t

$pbest(t)$ is the personal best value

$gbest$ is the global best value

$rand$ is the uniformly generated random number between 0 and 1

l_1, l_2 are learning factors.

ω is the inertia weight.

The inertia weight ω is calculated as

$$\omega = \omega_{max} - iteration * (\frac{\omega_{max}-\omega_{min}}{maxiter}) \quad (7)$$

where ω_{max} is the maximum inertia weight.

ω_{min} is the minimum inertia weight.

The pseudo code of PSO algorithm is given as follows:

```
Randomly initialize each particle's position in the population and its velocities
While (stopping criteria is not met)
{
    for each particle
         {
      Current value: Compute the fitness value.
      If the fitness value is better than the best fitness value pbest, then set pbest equal to the
      current value
      Select the particle with the best fitness value in the swarm and assign it as gbest.
    }
   for each particle
   {
       Compute particle velocity according to the equation (5).
      Update the position of the particle according to the equation (6).
    }
} // end while
Output the gbest solution.
```

4.1. Levy Flight

Levy flights (or Levy motion) [7-9] is a class of non-Gaussian random processes whose random walks are drawn from Levy stable distribution. This distribution is a simple power-law formula $L(s) \sim |s|^{-1-\beta}$ where $0 < \beta < 2$ is an index. Mathematically, a simple version of Levy distribution can be defined as:

$$L(s,\gamma,\mu)=\begin{cases}\sqrt{\frac{\gamma}{2\pi}}\ \exp\left[-\frac{\gamma}{2(s-\mu)}\right]\frac{1}{(s-\mu)^{\frac{3}{2}}} & \text{if } 0<\mu<s<\infty, \\ 0 & \text{if } s\leq 0\end{cases} \qquad (8)$$

where μ parameter is location or shift parameter, $\gamma > 0$ parameter is scale (controls the scale of distribution) parameter.

In general, Levy distribution should be defined in terms of Fourier transform.

$$F(k)=\exp[-\alpha|k|^{\beta}],\ 0<\beta\leq 2 \qquad (9)$$

where   is a parameter within [−1, 1] interval and known as skewness or scale factor. An index of o stability   (0, 2) is also referred to as Levy index. The analytic form of the integral is not known for general   except for a few special cases.

5. THE PROPOSED ALGORITHM

5.1Particle Swarm Optimization with Levy Flight (PSOLF)

In the proposed Particle Swarm Optimization (PSO) algorithm the updation of particles' position is done with the help of Levy Flight method. Levy flight performs a random walk in the search space and it provides high exploration. The new particle at the iteration is generated according to the equation (10):

$$x_i(t+1)=\mathrm{Levy\_walk}(x_i(t))+v_i(t+1) \qquad (10)$$

where $\mathrm{Levy\_walk}(x_i(t))$ is the levy flight function on the particle x_i.

$$Levy\_walk(X_i^{(t)}) = X_i^{(t)} + step \oplus random(size(X_i)) \qquad (11)$$

Where

$$step = stepsize * X_i^{(t)} \qquad (12)$$

and $stepsize$ is the value obtained from Levy flight method, $\oplus$ represents element-by-element multiplication.

5.2. Hybrid PSOLF-KHM Algorithm

Due to the disability of KHM algorithm in finding the global optima, we incorporate PSO with levy flight and KHM to structure a new hybrid algorithm named as PSOLF-KHM. As in [11] (Yang, Sun, & Zhang, 2009), PSOLF-KHM will also apply KHM with four iterations to all the particles in the swarm every eight iterations. In this proposed work, a particle is a k×d matrix of real numbers where the number of clusters is k and d is the number of features of data that is to be clustered and each row is the centroid of a cluster. The representation of a particle is shown in

Fig. 1. The objective function of the hybrid PSOLF-KHM is same as the objective function of the KHM algorithm.

s_{11}	s_{12}	s_{13}	s_{14}	…	s_{1d}	*Centre coordinates for cluster1*
s_{21}	s_{22}	s_{23}	s_{24}	…	s_{2d}	*Centre coordinates for cluster2*
s_{31}	s_{32}	s_{33}	s_{34}	…	s_{3d}	*Centre coordinates for cluster3*
s_{41}	s_{42}	s_{43}	s_{44}	…	s_{4d}	*Centre coordinates for cluster4*
⋮	⋮	⋮	⋮		⋮	⋮
s_{k1}	s_{k2}	s_{k3}	s_{k4}	…	s_{kd}	*Centre coordinates for cluster k*

Figure 1. Candidate solution representation

The summary of hybrid PSOLF-KHM algorithm is given as follows:

```
1.  Initialize the parameters population size $Pop_{size}$ , $\omega_{max}$,$\omega_{min,}$ learing factors $l_1$,$l_2$ and $N_{gen}$.
2.  Randomly initialize particles in the population of size $Pop_{size}$ .
3.  Assign G1=0.
4.  Assign G2=G3=0.
(PSOLF Method)
5.  While (G2<8)
        5.1 For each particle i do
                5.1.1  Update the particle's position and velocities according to equation (10)
                       and (5) respectively.
                5.1.2  Update pbest and gbest if necessary.
(KHM Method)
6.  For each particle i do
        6.1 Take the particle i position as initial cluster centers of the KHM algorithm.
        6.2 While (G3<4)
                6.2.1  Calculate KHM(X,C), membership and weight.
                6.2.2  Recompute cluster center$c_j$.
7.  G1=G1+1.If G1<$N_{gen}$ , go to step 4.
8.  Assign data point $x_i$ to cluster j with the biggest $m(c_j|x_i)$
```

1. EXPERIMENTAL RESULTS

All the three algorithms KHM algorithm, PSOKHM [11] and proposed PSOLF-KHM algorithm are implemented on an Intel Core i3-2350M 2.30 GHz using MATLAB 8.3, and seven test datasets are used to evaluate the performance of the proposed algorithm. Artificial datasets named Artset1 and Artset2 are drawn from Yang, Sun, & Zhang (2009) [11]. The remaining five datasets, namely, iris, thyroid, Contraceptive Method Choice (CMC), Crude oil and glass, are collected from ftp://ftp.ics.uci.edu/pub/machine-learning-databases/. The eight datasets used in this paper is described in Table 1. Table 2 lists the parameter settings of the PSOKHM and proposed algorithm PSOLF-KHM.

Table 1.Dataset Characteristics

Dataset Name	# of features	# of classes	# of instances(size of each class)
Artset1	2	3	300(100,100,100)
Artset2	3	3	300(100,100,100)
Iris	4	3	150(50,50,50)
Thyroid	5	3	215(150,35,30)
Cancer	9	2	683(444,239)
CMC	9	3	1473(629,333,511)
Glass	9	6	214(70,17,76,13,9,29)

TABLE 2. PARAMETER SETTINGS FOR PSOKHM AND PSOLF-KHM

Parameter	Value
Iteration Count (N_{gen})	5
Population Size(Pop_{size})	18
ω_{max}	0.9
ω_{min}	0.4
l_1	1.49618
l_2	1.49618

6.1. Data Sets

The seven datasets used in this paper is described as follows:
Let,
N is the total number data objects to be clustered,
M is the number of attributes for each data object and
K is the number of clusters to be partitioned to.

Data set 1: Artificial dataset 1 (Artset1)
(N=300, M=2, K=3)

This dataset containing two featured problem is drawn from three independent bivariate normal distributions of three classes, where classes are distributed according to $N2\left(\mu = \begin{pmatrix}\mu_{i1}\\ \mu_{i2}\end{pmatrix}, \Sigma = \begin{bmatrix}0.4 & 0.04\\ 0.04 & 0.4\end{bmatrix}\right)$, i=1,2,3, $\mu_{11} = \mu_{12}$ =-2, $\mu_{21} = \mu_{22}$ =2, $\mu_{31} = \mu_{32} = 6$, μ and $\sum$ being mean vector and covariance matrix respectively. The data set is shown in Fig. 2.

Data set 2: Artificial dataset 1 (Artset2)

(N=300, M=3, K=3)

This artificial dataset contains samples drawn from five independent uniform distributions with ranges of [10, 25], [25, 40], and [40, 55]. Each class contains 100 samples, totally 300 samples of 3 classes and each sample has 3 attributes. The data set is shown in Fig. 3.

Data set 3: Iris data

(N=150, M=4, K=3)

This dataset contains three categories of 50 objects each, where each category refers to a type of iris plant. Totally there are 150 instances with three attributes, which are sepal length in cm, sepal width in cm, petal length in cm and petal width in cm.

Data set 4: Thyroid gland data

(N=215, M=5, K=3)

This dataset contains three categories of human thyroid diseases, namely, normal, hypothyroidism and hyperthyroidism. In the thyroid gland dataset, there are 215 samples with total of six attributes. The first is a class attributes indicating 1 for normal, 2 for hyperthyroidism and 3 for hypothyroidism. The remaining five attributes are considered for clustering the data, namely the T3-resin uptake test, total Serum thyroxin as measured by the isotopic displacement method, total serum triiodothyronine as measured by radioimmuno assay, basal thyroid-stimulating hormone as measured by radioimmuno of 200 mg of thyrotropin releasing-hormone and the basal value. All attributes are continuous.

Data set 5: Cancer data

(N=683, M=9, K=2)

This data set contains 683 data objects that are categorized into two categories: malignant (444 objects) and benign (239 objects). Each data object is featured by nine features: clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, bland chromatin, normal nucleoli, single epithelial cell size, bare nuclei and mitoses.

Data set 6: Contraceptive Method Choice (CMC) data

(N=1473, M=9, K=3)

This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The data objects about the married women who were either not pregnant or not aware if they were at the time of interview. The problem involves predicting the choice of the current contraceptive method of a woman based on her socio-economic and demographic characteristics. This dataset contains 1473 objects with nine attributes and three clusters.

Data set 7: Glass data

(N=214, M=9, K=6)

This dataset contains 214 objects with nine attributes, namely, refractive index, sodium, magnesium, potassium, calcium, aluminum, silicon, barium and iron. The data were sampled from six different types of glass: float processed building windows, non-float processed building windows, float-processed vehicle windows, containers, tableware and headlamps.

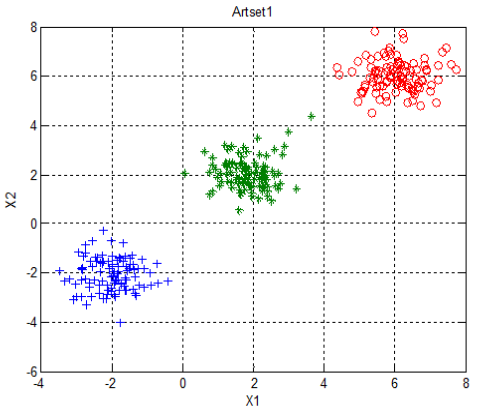

Figure 2. Artificial Dataset1 (Artset1)

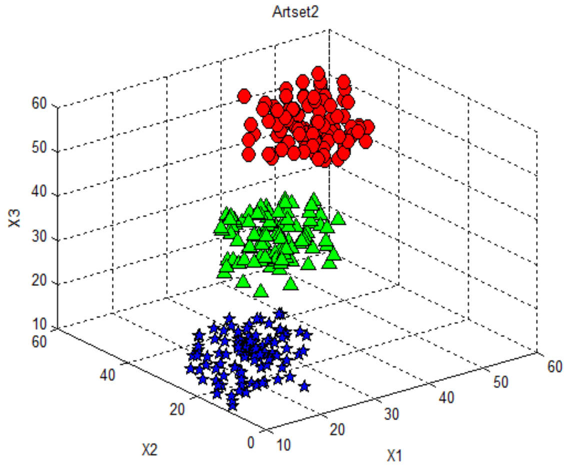

Figure 3. Artificial Dataset2 (Artset2)

6.2. Performance Evaluation

The quality of clustering algorithms is measured using objective function value and F-measure. The smaller the objective function value is, the quality of clustering will be higher.

The F-measure makes use of the ideas of precision and recall values used in information retrieval. The *precision P(i,j)* and *recall R(i,j)* for each class i of each cluster j are calculated as

$$P(i,j)=\frac{\gamma_{ij}}{\gamma_j} \tag{11}$$

$$R(i,j)=\frac{\gamma_{ij}}{\gamma_i} \tag{12}$$

where,

γ_i : is the number of members of class i
γ_j : is the number of members of cluster j
γ_{ij}: is the number of members of class i in cluster j

The corresponding *F-measure F(i,j)* is given in Eq. (13):

$$F(i,j)=\frac{2*P(i,j)*R(i,j)}{P(i,j)+R(i,j)} \tag{13}$$

Then the definition of *F-measure* of a class i is given as

$$F_{tot}=\sum_i \frac{\gamma_i}{n} \max_j (F(i,j)) \tag{14}$$

where, n is the total number of data objects in the collection. In general, the larger the F-measure gives the better clustering result.

6.3. Results Discussion

In this paper, to compare the performance of proposed algorithm, each algorithm has been run for 10 times. The average and the standard deviation of each algorithms' objective function values and the F-measure values (over 10 runs) are given in table 3-5 for p=2.5,3,3.5 respectively. Since p is a key parameter to get good objective function values, each algorithm is run with different p values given above. Also tables 3-5 include the average runtime of the algorithms. The quality of clustering is evaluated using KHM(X, C) and the F-Measure. Runtimes (s) are additionally provided. The table shows means and standard deviations (in brackets) for 10 independent runs. Bold face indicates the best result out of the three algorithms.

When p=2.5, PSOKHM produces better average KHM value and F-measure value for iris dataset than other two algorithms and for other datasets such as Thyroid, Glass, Cancer, PSOLF-KHM gives better results than other two algorithms. When p=3, 3.5, PSOLF-KHM performs best in case of average KHM(X, C) and F-measures.

The average of objective function value for PSOLF-KHM algorithm is better than KHM and PSOKHM algorithms. This means that the proposed PSOLF-KHM performs with very high exploration in the search space before converging to optima.

Table 3. RESULTS OF KHM, PSOKHM, AND PSOLF-KHM CLUSTERING WHEN P = 2.5

	KHM	PSOKHM	PSOLF-KHM
ArtSet1			
KHM (X,C)	670.046 (0.002)	670.033 (0.004)	**670.032 (0.004)**
F-Measure	0.997 (0.000)	0.997 (0.000)	0.997(0.000)
Runtime	0.064 (0.012)	5.900(0.661)	6.244(0.928)
ArtSet2			
KHM (X,C)	107619.088 (0.004)	107615.104 (1.862)	**107601.814 (5.104)**
F-Measure	1.000(0.000)	1.000(0.000)	1.000(0.000)
Runtime	0.079(0.014)	6.125(0.276)	6.374(0.938)

Iris			
KHM (X,C)	148.957 (0.042)	**148.876 (0.013)**	148.895 (0.009)
F-Measure	0.885(0.002)	0.887(0.003)	0.885(0.000)
Runtime	0.046(0.009)	3.440(0.090)	3.372(0.124)
Thyroid			
KHM (X,C)	221503.499 (112.609)	220815.727 (133.677)	**220624.245 (117.106)**
F-Measure	0.776(0.051)	0.780(0.048)	0.780(0.050)
Runtime	0.196(0.024)	4.732(0.138)	4.937(0.601)
Cancer			
KHM (X,C)	57169.069(0.165)	57044.633(28.068)	**56885.391(27.074)**
F-Measure	0.624(0.009)	0.627(0.008)	0.627(0.007)
Runtime	0.133(0.021)	11.625 (0.638)	11.300(0.903)
CMC			
KHM (X,C)	96200.501 (3.832)	96186.522 (6.505)	**96147.430 (27.550)**
F-Measure	0.418(0.001)	0.419(0.001)	0.419(0.001)
Runtime	0.644(0.170)	36.236(1.820)	37.053(0.372)
Glass			
KHM (X,C)	1210.944 (17.347)	1177.108 (25.285)	**1175.117 (32.654)**
F-Measure	0.431(0.116)	0.449(0.064)	0.451(0.094)
Runtime	0.199(0.033)	10.825(0.797)	10.591 (0.726)

TABLE 4. RESULTS OF KHM, PSOKHM, AND PSOLF-KHM CLUSTERING WHEN P = 3

	KHM	PSOKHM	PSOLF-KHM
ArtSet1			
KHM (X,C)	701.628(0.047)	701.570 0.039)	**701.565 (0.029)**
F-Measure	0.999(0.001)	0.999(0.002)	1.000(0.001)
Runtime	0.080(0.019)	5.740(0.289)	5.758(0.286)
ArtSet2			
KHM (X,C)	266954.835 (0.107)	266897.131(40.087)	**266802.473 (33.659)**
F-Measure	1.000(0.000)	1.000 (0.000)	1.000 (0.000)
Runtime	0.107(0.013)	6.275(0.408)	7.271(1.005)
Iris			
KHM (X,C)	126.103(0.012)	125.974 (0.060)	**125.97 (0.072)**

F-Measure	0.888(0.003)	0.890(0.003)	0.891(0.002)
Runtime	0.038(0.007)	3.157(0.131)	3.377(0.114)
Thyroid			
KHM (X,C)	978309.547 (2416.467)	958946.937 (2784.067)	**957101.217 (3352.464)**
F-Measure	0.763(0.049)	0.764(0.053)	0.764(0.082)
Runtime	0.288(0.095)	4.685(0.219)	4.779(0.325)
Cancer			
KHM (X,C)	113726.852 (0.057)	112780.008(251.681)	**111711.103 (277.841)**
F-Measure	0.960(0.004)	0.960(0.004)	0.961(0.004)
Runtime	0.151(0.021)	11.625 (0.322)	11.498 (0.179)
CMC			
KHM (X,C)	187015.903 (6.244)	186912.408(54.517)	**186620.893 (85.305)**
F-Measure	0.418(0.001)	0.419(0.002)	0.419(0.002)
Runtime	0.670(0.083)	38.302(4.681)	35.827(1.627)
Glass			
KHM (X,C)	1514.756 (184.021)	1395.547 (0.472)	**1394.019 (1.365)**
F-Measure	0.441 (0.101)	0.454 (0.112)	0.487 (0.048)
Runtime	0.244 (0.073)	10.452 (0.254)	12.146 (0.596)

Table 5. RESULTS OF KHM, PSOKHM, AND PSOLF-KHM CLUSTERING WHEN P = 3.5

	KHM	PSOKHM	PSOLF-KHM
ArtSet1			
KHM(X,C)	763.641 (1.166)	762.328 (0.152)	**762.138 (0.082)**
F-Measure	1.000(0.000)	1.000(0.000)	1.000(0.000)
Runtime	0.090 (0.016)	6.039(0.359)	5.951 (0.341)
ArtSet2			
KHM(X,C)	663971.899 (0.282)	663532.961 (278.892)	**663113.279 (185.425)**
F-Measure	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Runtime	0.159(0.014)	5.962(0.457)	6.234(0.407)
Iris			
KHM(X,C)	110.108 (0.227)	109.606 (0.132)	**109.527 (0.207)**

F-Measure	0.891(0.000)	0.891(0.002)	0.892(0.000)
Runtime	0.041(0.009)	3.188(0.285)	3.152(0.212)
Thyroid			
KHM(X,C)	4260105.456 (32.357)	4156426.270(14210.111)	**4130234.327 (20993.978)**
F-Measure	0.714(0.042)	0.714(0.098)	0.715(0.090)
Runtime	0.449(0.065)	4.712(0.307)	4.775(0.440)
Cancer			
KHM(X,C)	232191.308 (0.348)	227942.780 (1039.816)	**223482.296 (1024.472)**
F-Measure	0.961 (0.003)	0.961 (0.004)	0.961 (0.004)
Runtime	0.221(0.024)	10.753 (0.131)	11.245 (0.542)
CMC			
KHM(X,C)	380724.334 (11.997)	380322.502 (233.470)	**379609.696 (469.455)**
F-Measure	0.419(0.001)	0.418(0.002)	0.419(0.002)
Runtime	1.507(0.436)	34.594(0.821)	35.558(0.906)
Glass			
KHM(X,C)	1871.285(41.652)	1852.289(3.473)	**1834.375 (7.149)**
F-Measure	0.427(0.093)	0.436(0.101)	0.475(0.059)
Runtime	0.217(0.033)	10.647(0.360)	10.965(0.510)

7. CONCLUSION

This paper presents the hybrid clustering algorithm based on PSO with Levy Flight and KHM algorithms (PSOLF-KHM). This algorithm is tested on seven datasets. Experimental results show that the PSOLF-KHM algorithm is better than PSOKHM and KHM algorithm in terms of objective function value. The algorithm puts together the merits of KHM, PSO and Levy flight so that it searches the cluster center efficiently and thus achieves global optima.

However, PSOLF-KHM algorithm requires more time than KHM to run. This is the one limitation of PSOLF-KHM. In future, KHM with other stochastic optimization techniques will be used to solve global optimization problems.

REFERENCES

[1] P.N. Tan, M.Steinbach, and V.Kumar. 2005. *Introduction to data mining*. Boston: Addison-Wesley.pp. 487–559.

[2] Jiawei han and Michelin Kamber. 2010. *Data mining concepts and techniques*. Elsevier.

[3] B.Zhang, Hsu, and Dayal, K-harmonic means – a data clustering algorithm. Technical Report HPL-1999-124. Hewlett-Packard Laboratories.

[4] B.Zhang, M.Hsu, and U.Dayal, K-harmonic means. In: International workshop on temporal, spatial and spatio-temporal data mining, TSDM2000. Lyon, France,September 2000.

[5] J.Kennedy, and R.C.Eberhart . 1995. Particle Swarm Optimization. Proceedings of the IEEE International Joint Conference on Neural [1] Networks, 4:1942–1948.

[6] D.W.Van der Merwe. and A.P. Engelbrecht. 2003. Data clustering using particle swarm optimization. Proceedings of IEEE Congress on Evolutionary Computation, 1:215-220.

[7] A.V.Chechkin, ,R. MetzlEr, J. Klafter, and V.Y.Gonchar. 2008. Introduction to the theory of Lévy flights. in R. Klages, G. Radons, I.M. Sokolov (eds.), Anomalous Transport: Foundations and Applications, John Wiley & Sons, pp.129–162.

[8] X.S. Yang., 2010. Engineering Optimization an Introduction with Metaheuristic Applications. first ed., John Wiley & Sons, New Jersey.

[9] Huseyin Haklı and Harun U guz .() . 2014. A novel particle swarm optimization algorithm with Levy flight. *Applied Soft Computing*,23:333–345.

[10] Y. T .Kao, E.Zahara, and I.W. Kao. 2008. A hybridized approach to data clustering. *Expert Systems with Applications*, 34:1754–1762.

[11] F. Q Yang, T.L.Sun, and C.H. Zhang. 2009. Efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization. *Expert Systems with Applications*, 36:9847–9852.

[12] Hua Jiang, Shenghe Yi, Jing Li, Fengqin Yang and Xin Hu. 2010.Ant clustering algorithm with K-harmonic means clustering. *Expert Systems with Applications*, 37: 8679–8684.

[13] Minghao Yin, H.Yanmei , Fengqin Yang, Xiangtao Li and Wenxiang Gu. 2011. A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering. *Expert Systems with Applications*, 38:9319–9324.

[14] J.Kennedy, and R.C.Eberhart. 2001. Swarm Intelligence. Morgan Kaufmann.

[15] Zulal Gungor and Alper Unler. 2008. K-Harmonic means data clustering with tabu-search method. *Applied Mathematical Modelling*, 32:1115–1125.

[16] Z.Güngör, and A. Ünler. 2007. K-harmonic means data clustering with simulated annealing heuristic. *Applied Mathematics and Computation*, 184:199–209.

[17] R.Jensi,and G.Wiselin Jiji. 2013. A Survey on optimization approaches to text document clustering. *International Journal on Computational Science and Applications*, 3: 31-44.

[18] R.Jensi,and G.Wiselin Jiji. (2016), “An enhanced particle swarm optimization with levy flight for global optimization” Applied Soft Computing, Vol.43, pp. 248–261.