Original Article

Print ISSN: 2321-6379
Online ISSN: 2321-595X
DOI: 10.17354/ijss/2017/33

K-means Optimization Clustering Algorithm Based on Hybrid PSO/GA Optimization and CS validity index

K Jahanbin[1]*, F Rahmanian[2], H Rezaie[3], Y Farhang[4], A Afroozeh[5]

[1]Department of Computer, Kerman Branch, Islamic Azad University, Kerman, Iran, [2]Department of Computer, Kerman Branch, Islamic Azad University, Kerman, Iran, [3]Department of Managment, Firozkuh Branch, Islamic Azad University, Friozkuh, Iran, [4]Faculty of Computer, Khoy Branch, Islamic Azad University, Khoy, Iran, [5]University of Larestan, Lar, Iran

Abstract

K-means is the most usable clustering partition-based algorithm, however, this algorithm is highly dependent on the initial mass centers and the selection of the number of clusters to start; hence it is rapidly trapped in the local optimum. A new framework has been presented in this paper based on the combination of Genetic, Particle swarm Optimization algorithms and taking advantage of CS validation index as the cost function called Hybrid-PSO-GA. At first, in the proposed method the location and number of clusters have been found by the HGA algorithm and then the cost function for each iteration is calculated, finally, the optimal solution with the best value of cost function, location and number of mass centers is sent to the K-means algorithm. The proposed approach in addition to solve the problem of random start and falling K-means algorithm into the local optimum, because of using both evolutionary and swarm intelligence approaches in comparison to the optimization method of K-means simply with genetics or PSO in terms of final clustering, faster and more accurate convergence to global optimum and the number of function evaluation has better performance and can achieve the most effective results of clustering.

Keywords: K-means, Hybrid PSO/GA Optimization, Clustering

INTRODUCTION

Clustering is one of the most important issues of data mining, unlike classification, clustering is an unsupervised learning method and its ultimate goal is to maximize the separation and minimize the cohesion. Hence the clustering can also be regarded as an optimization problem and the evolutionary and swarm intelligence methods can be used to solve and optimize the clustering criteria. In general, clustering methods can be divided into fourgroups:

- Partition-Based Methods
- Hierarchical methods
- Grid-Based Methods
- Model-based methods

Access this article online

IJSS

www.ijss-sn.com

Month of Submission : 06-2017
Month of Peer Review : 06-2017
Month of Acceptance : 07-2017
Month of Publishing : 07-2017

K-means based is a partition-based method and due to the simple structure, fast implementation, and rapid convergence rate has numerous applications in the field of Data-mining & Machine Learning including: Clustering, Image Segmentation, and Pattern recognition; However Partition-based clustering algorithms such as K-means, K-medoids and Fuzzy K-Means (FCM) have big problems such as extreme sensitivity to initial starting location and are fast convergence to local optimum [1].

In this paper, a hybrid approach based on PSO and GA algorithms called Hybrid PSO/GA (HPG) has been used to deal with the problems of the K-means algorithm and other Partition-Based Clustering algorithms. Due to the high power of genetic algorithms in searching the answer it is an appropriate item to optimize the objective function; however, evolutionary algorithms have only a mere optimization approach, while swarm intelligence algorithms like PSO produce the optimal solution based on the experiences and observations of each particle, thus, in this paper, the combination of both methods have been used to exploit their benefits aimed at finding an effective method of clustering.

Corresponding Author: K. Jahanbin. Department of Computer, Kerman Branch, Islamic Azad University, Kerman, Iran.
E-mail: Mohagheghnia7877@gmail.com

In the proposed approach, K of the cluster center with the minimum solution in terms of cost function CS produced by HPG algorithm and is sent to K-means algorithm or any other Partition-Based algorithm; Then the algorithm based on the position of the centers of clusters has conducted the clustering and the results are analyzed with the criteria Separation, Cohesion, NFE and value of cost function CS per iteration, the analysis results show that the algorithm HPG solves the problems of random start and falling into the local optimum of K-means and also, uses the simple structure, fast implementation and rapid convergence of Partition-Based algorithms.

A REVIEW OF PREVIOUS LITERATURE

Lu et al. provided fast genetic-based K-means algorithm (FGKA) and genetic growing algorithm inspired by Christa and Moretti algorithm called GKA[2], features of these two algorithms are higher speed and better convergence to global optimum [3, 4].

In [5] an automatic clustering method based on Differential Evolution for the classification of large data without the knowledge of their labeling is provided, Automatic segmentation image is also used by the author in the algorithm; According to the author, the proposed method has appropriate speed, stability, and convergence to the global optimum.

In[1] an improved K-means technique called PM-Kmeans has been provided that, sends the optimal centers of clusters to K-means using PSO and K-means has conducted the clustering process and then two solutions are combined.

As can be seen, all methods are based solely on evolutionary algorithms or swarm intelligence, but in the proposed HPG algorithm, advantages of both methods and CS cost function have been used in automatic clustering.

SCIENTIFIC BACKGROUNDS

K-means Algorithm

K-means [6] is a Vector Quantization method widely used in Cluster Analysis and Data Mining. K-means isan NP-Hard problem, though it can be improved by Meta-Heuristic algorithms and converged towards a local optimum rapidly. K-means got a set of observations $\vec{X} = \{x_1, x_2, \ldots, x_n\}$, $\vec{X} \in \mathbb{R}^d$ and assigned it into K clusters of $\vec{S} = \{S_1, S_2, \ldots S_k\}$ where, $K << N$. The aim of K-means is to minimize the Within-Cluster Sum of Square Distance (WCSS), in other words:

(1) $\arg\min_s \sum_{i=0}^{K} \sum_{x\in S_i} \| x_i - s_i \|^2$.

PSO Algorithm

PSO [7] is essentially a continuous algorithm in the field of Swarm Intelligence (SI). From another perspective, due to the implementation of an optimization process in each run round and ability to obtain the solution with minimum information, PSO can be considered as an evolutionary algorithm (EA) and Meta-Heuristic algorithm.

The PSO algorithm works by having a population (called a swarm) of candidate solutions (called particles). these particlesmove around the search-space with velocity determined by the best positionexperienced by each particle. Particle saves its best-experiencedposition displayed by $x^{i,best}$ in its memory and compares it with the best position experienced by all particles shown with x^{Gbest} and, all particles set their new position according to the values of $x^{i,best}$ and x^{Gbest} to reach new solution.

If x^i, v^i and $x^{i,best}$ are the position, velocity and the best position experienced by *particle*$_i$, respectively, ω is inertia coefficient and x^{Gbest} is best position for the accumulation of particles, equations describing the behavior of the particles can be written as follows:

2) $v^i[t+1] = \omega r^i[t]$

$+c_1 r_1 \left(x^{i,best}[t] - x^i[t] \right)$

$+c_2 r_2 \left(x^{Gbest}[t] - x^i[t] \right)$

$x^i[t+1] = x^i[t] + v^i[t+1]$

Where $x^i[t]$ is the current position in t-thtime (or iteration), $v^i[t+1]$ is the new velocity at time t + 1'th and $x^i[t+1]$ is the new position of the particle in time t + 1'th, $r_1, r_2 \sim \mathcal{U}(0,1)$ and c_1, c_2 are respectivelythe factor of personal and collective learning set according Constriction coefficients property.

Constriction Coefficients[8]: Consider to constants $\phi_1, \phi_2 > 0$, we define:

3) $\phi_1, \phi_2 > 0 \Rightarrow \phi \triangleq \phi_1 + \phi_2 > 4$

Now the number of ξ is as follows:

4) $\xi = \dfrac{2}{\phi - 2 + \sqrt{\phi^2 - 4\phi}}$

We define the coefficients:

5) $\begin{cases} \omega = \xi \\ C_1 = \xi\phi_1 \\ C_1 = \xi\phi_2 \end{cases}$

Proven at optimal state of $\phi_1 = \phi_2 = 2.05$, ω = 0.7298 and $C_1 + C_2 = 1.4962$.

Genetic Algorithm

Genetic Algorithm[9] is an optimization and randomizes search technique, inspired by the principles of evolution and natural genetics. GA can search in complex, large and multidimensional spaces to find aNear-optimal solution for the objective function in an optimization problem.

In GA, the parameters are encoded in the search space in a string called chromosomes, and a set of chromosomes have formed a population. In general, either binary or continuously, genetic algorithm can be written as follows:

```
1: t⟵0
2: Initialize Population [P(t)];
3: Evaluates The Population [P(t)];
4: While (not termination) do
5: $P'$ = Variation [$P'$(t)]; {Create of new solutions}
6: Evaluate Populations [$P'$ (t)]; {Evaluates the new
    solutions}
7: P(t+1) ⟵ Apply Genetic Operation [$P'$ (t)∪ Q]; { New
generation Pop}

8: t⟵t+1;

9: End While.
```

Mutation & Arithmetic Crossover: If n_{var} indicates the number of unknowns, and $\vec{X} = \{x_1, x_2, \ldots, x_{n_{var}}\}$, $\vec{X}_i \in [x_{min}, x_{max}]$ represents the population, then each parent is defined as follows (two parents are shown):

$$Parents_{1,2} \begin{cases} X_1 = \{x_1, x_2, \ldots, x_{1n}\} \\ X_2 = \{x_1, x_2, \ldots, x_{2n}\} \end{cases}$$

$$(mask)\alpha = \{x_1, x_2, \ldots, x_n\}$$

$$Offspring_{1,2} \begin{cases} Y_{1i} = \alpha_i x_{1i} + (1-\alpha_i) x_{2i} \\ Y_{2i} = \alpha_i x_{2i} + (1-\alpha_i) x_{1i} \end{cases}$$

α is a parameter for themutation but this parameter α not alone enough, because offspring in the best conditions are similar to one of their parents and are not able to search in the space further than its parents. If offspring are in the range 0≤α≤1 in the best case, they will be similar to one of their parents but if the range with minor changes will be $-\gamma \le \alpha \le 1+\gamma$, offspring can obtain more optimal solution than their parent (for example γ =0.05).

CS Clustering Validity Index

CSclustering validity index[10] is recently proposed to assess the validity and objective function is clustering algorithms proposed. Before applying CS measuring method, cluster centers are calculated by calculating the average of data vectors of each cluster center:

$$6)\ m_c^{\rightarrow} = {}^{1}\!/_{N_i} \sum_{x_j \in C_i} x_j^{\rightarrow}$$

The measure of the distance between points $x_i^{\rightarrow}$ and $x_j^{\rightarrow}$ is shown by $d\left(x_i^{\rightarrow}, x_j^{\rightarrow}\right)$. Therefore, the CS measure can be defined as the following:

$$7)\ CS(k) = \frac{{}^{1}\!/_{k} \sum_{i=1}^{k} [\ {}^{1}\!/_{N_i} \sum_{X_i^{\rightarrow} \in C_i} \max\limits_{X_q^{\rightarrow} \in C_i} \left\{ d\left(X_i^{\rightarrow}, X_q^{\rightarrow}\right) \right\}\ ]}{{}^{1}\!/_{k} \sum_{i=1}^{k} [\ \underset{j \in k\ ,\ j \neq i}{mix} \left\{ d\left(m_i^{\rightarrow}, m_q^{\rightarrow}\right) \right\}]}$$

$$= \frac{\sum_{i=1}^{k} [\ {}^{1}\!/_{N_i} \sum_{X_i^{\rightarrow} \in C_i} \underset{X_q^{\rightarrow} \in C_i}{max} \left\{ d\left(X_i^{\rightarrow}, X_q^{\rightarrow}\right) \right\}\ ]}{\sum_{i=1}^{k} [\ \underset{j \in k\ ,\ j \neq i}{mix} \left\{ d\left(m_i^{\rightarrow}, m_q^{\rightarrow}\right) \right\}]}$$

It is observed that most CS divided the longest distance between cluster members over the shortest distance between the clusters, consequently, CS investigates two factors of Cohesion and Separation which are the characteristics of a suitable clustering algorithm.

Exception: the error of divide by zero may occur in CS calculations, this error will be occurred when one of the centers of clusters is outside the boundary of distributions of the dataset.To avoid this problem, if the cluster center was without data or had very small number of data (for example, 2 or 3 data), this cluster center is penalized by assigning large values:

$$8)\ 10 * norm \,|\left| Max\left(X^{\rightarrow}\right), \square Min\left(X^{\rightarrow}\right) \right||^2$$

HYBRID PSO/GA METHOD

PSO algorithm often has a strong tendency to fall into local optimum because there is swarm in the lack diversity.On the other hand,genetic algorithm is only a mere optimizer algorithm and does not use the properties of swarm intelligence that is, the relationship between particles and keeping the best memories, hence, HPG algorithm tries to take advantage of the properties of both evolutionary and swarm intelligence approaches.

As mentioned above, the cost function CS clearly supports the cohesion and separation criteria, hence the PSO in the HPG algorithm finds the amount of globalbest at first,then genetic algorithm operators such as Crossover and Mutation are applied on the particles; the goal is to expand the search space and achieve more optimal answers.

Particle Population Encoding and Clustering Cost

HPG algorithm answers are entitled as particle population, the objective function has been designed in accordance with [5] and clustering validity index CS is attached to it.

Particle population encoding: Suppose the dataset is equal to $\vec{X} = \{x_1, x_2, \ldots, x_n\}, \vec{X} \in \mathbb{R}^{n*d}$, the cluster centers are represented with $\vec{M} = \{m_1, m_2, \ldots, m_k\}, \vec{X} \in \mathbb{R}^{k*d}, K << N$, and to activate the center of each cluster the activation matrix of $\vec{A} = \{a_1, a_2, \ldots, a_k\}, \vec{X} \in \mathbb{R}^k$ is used, then the particle population matrix is $k*(d+1)$ dimensional and in the following encoding form:

$$K\left\{\begin{bmatrix} m_1 & \cdots & a_1 \\ m_2 & & a_2 \\ \vdots & \ddots & \vdots \\ m_k & \cdots & a_k \end{bmatrix}\right. \\ \underbrace{\hphantom{m_1 \cdots}}_{\substack{\text{Cluster}\\ \text{Centriods}\\ \text{d(dim)}}} \underbrace{\hphantom{a_k}}_{\substack{\text{Activation}\\ \text{(one column)}}}$$

We define $\vec{X} = Max(\vec{X})$ and $\vec{X} = Min(\vec{X})$ then, $\vec{X} \le m_i \le \vec{X}, m_i \in \mathbb{R}^{k*d}$ and $a_i \sim \mathcal{U}(0,1), a_i \in \mathbb{R}^k$, if , is a_ i≥0.5 the cluster center is active. The following program shows the cost function that uses the CS index to validate the cluster centers:

```
Function Clustering_Cost ( $\vec{X}$ )
{
    $\vec{M} = \vec{X}$ (:, 1:end-1) % All columns except the last column
    $\vec{A} = \vec{X}$ (:,end) % the last column
    $\vec{A}$ = CS_index( $\vec{M}$ , $\vec{X}$ )
    IF $\left[\sum_i Numel(a_i \ge 0.5)\right]$ < Two Cluster Centers
        find two biggest $a_i \xrightarrow[then]{}$ Active
    Else
        $\forall_{a_i} \ge 0.5 \xrightarrow[then]{}$ Active
    Output:
    Cluster_Cost. $\vec{A}$
}

Function CS_index ( $\vec{X}$ , $\vec{M}$ )
{
    $d_p^{max} = max_{x_p, x_q \in C_i} d(x_p \ x_q)$ % The maximum of Pairwise distance between the two clusters.
    IF $\left[\sum_i Numel(C_i)\right]$ > a few number of Cluster
        $\hat{d}_i = 1/N_i \sum_{x_p \in C_i} d_p^{max}$
    Else
        $\hat{d}_i = 10 * norm\,|\left|Max(X^{\rightarrow}), Min(X^{\rightarrow})\right||^2$
    $CS = \frac{\sum_i \hat{d}_i}{\sum_i \min_{j \neq i} d(m_i, m_j)}$ % min Pairwise distance between two cluster
    Output:
    CS_index. $\vec{M}$
    CS_index.CS
}
```

One of the benefits of HPG method is replacement of CS function by any other clustering validity index, now we can consider the cost function as follows:

Clustring_Cost=CS

And fitness function value is calculated as follows:

$$Fitness = \frac{1}{CS + \varepsilon}$$

HPG Algorithm

As mentioned, HPG algorithm is a combination of PSO and GA algorithms, in HPG velocity parameter of PSO algorithm has been applied to the GA algorithm population in the parts of cross-Over and mutation so that, GA algorithm can search more space to increase the exploration;On the other hand, by adding GA parameters such as sorting the population, Cross-Over, and mutation on PSO particles, more accurate optimization will be occurred which is on of the tasks of evolutionary algorithms and will lead to increased exploitation.

In addition to the benefits of the evolutionary algorithms and swarm intelligence, the proposed method deliver the final clustering to the K-means algorithm due to higher execution speed, on the other hand, HPG algorithm has solved the problem of falling of PSO into local optimization and eliminates the severe dependence of partition-Based methods such as K-means to the initial starting centers.

In all the algorithms it is assumed that $sp_i.best$ is the best local value of the objective function related to *swarm - population*$_i$ and sp^{Gbest} is the best value obtained by the total population of particles to the i-th round. $\vec{M}$ also indicates the position and the number of all clusters and $\vec{X}$ represents the tested dataset.

```
Algorithm 1: Automatic PSO Clustering
Input:
Dataset: $\vec{X} = \{x_1, x_2, \ldots, x_n\}$
Cluster Centroid: $\vec{M} = \{m_1, m_2, \ldots, m_k\}$
Max-Sub-Iteration-PSO= 1
Sp_population =100;
Output: Sp_best.cost={ $sp_1.best$ , $sp_2.best$ ,…,
$sp_{Max-iteration}.best$ };
$sp^{Gbest}$ ;
Procedure Automatic PSO clustering
    assign $\omega, C_1, C_2$ using Equation(5);
    For i=1 to Max-Sub-Iteration-PSO
        For i=1 to Sp_population
            UPDATE Velocities for $sp_i$ using Equation(2);
            min ( $sp_i.velocity$ )≤ $sp_i.velocity$ ≤ max ( $sp_i.velocity$ ); % Apply Velocity Limit
            UPDATE Position for $sp_i$ using Equation(2);
            IF ( $sp_i.position$ < min ( $\vec{X}$ ) || $sp_i.position$ > max ( $\vec{X}$ )) % Velocity Mirror Effect
                $sp_i.position = -sp_i.position$ ;
            END IF
            [$sp_i$. cost, $sp_i$. index]= Clustring_Cost( $sp_i$ );
            IF ( $sp_i.best$ < Clustring_Cost( $sp_i$ )) % Update $sp_i.best$
                UPDATE $sp_i.cost$ , $sp_i.position$ ;
            END IF
            IF ( $sp^{Gbest} < sp_i.best$ ) % Update $sp^{Gbest}$
                $sp^{Gbest} = sp_i.best$ ;
            END IF
        END FOR
    END FOR
END Procedure
```

Automatic GA algorithm is as follows:

```
Procedure Automatic PSO clustering
    $n_{crossover} = 2\frac{P_{crossover} * swarm - popluatio}{2}$ ;
    For i=1 to Max-Sub-Iteration-GA
        For i=1 to $n_{crossover}$
            $Parent_{i_1}$ = RouletteWheelSelection ( $e^{-\beta C_i}$ );
            $Parent_{i_2}$ = RouletteWheelSelection ( $e^{-\beta C_i}$ );
            UPDATE $sp_i.best$ , $sp_i.velocity$ ;
        END FOR
        Merge & Sort population;
        IF ( $sp_i.best < sp_i.best.cost$ )
            $sp_i.best.cost = sp_i.best$ ;
        END IF
        IF ( $sp^{Gbest} < sp_i.best.cost$ )
            $sp^{Gbest} = sp_i.best.\cos ta$ ;
        END IF
    END FOR
END Procedure
```

Termination condition: HPG algorithm has two combinational sub-programs of PSO and GA, in each iteration of the algorithm once PSO and once GA are executed;It can be concluded if global best does not change after 10 runs, the algorithm reached a stable level and algorithm does not have to be continued more frequently (this result is clearly seen in the Result part of figures 3 and 6).

Termination condition is as follows:

```
IF (iteration >=30)
    IF ($sp^{Gbest}$ (i) == $sp^{Gbest}$ (i-10))
        Break;
    END IF
END IF
```

Two algorithms of Automatic PSO and GA clustering and termination condition were defined above, the main body of HPG algorithm is as follows:

Algorithm 3: Automatic HPG Clustering

```
Procedure Automatic HPG Clustering
For i=1 to Max-iteration
    Automatic PSO Clustering;
    Automatic GA Clustering;
END FOR
$\vec{M}$ = Numel ($sp^{Gbest}$ .m);
Best_Solution = ($sp^{Gbest}$ .cost);
K-means ($\vec{M}$ , $\vec{X}$) % K-medoids or Fuzzy-K-means
{ Repeat Until Convergence
    For $\forall_i$
        $C^i = \arg\min_J \| x^j - C^j \|_2$ ;
    For each j
        $C^j = \frac{\sum_{i=1}^{C} 1\{C^i = j\} x^i}{\sum_{i=1}^{M} 1\{C^i = j\}}$ ;
}
END Procedure
```

RESULTS

Two liner and non-Liner standard dataset named as $X_1^{\to}$ and $X_2^{\to}$ have been used for the analysis. $X_1^{\to}$ is associated with 1000 samples and five cluster centers in positions [-2 0], [3, 5], [9 1], [-2 10] and [8, 10]. $X_2^{\to}$ is a non-linear dataset containing 9000 samples and 3 cluster centers in positions [0 0], [3 3] and [3 -3], Nonlinear dataset has been tested because of the weakness of Partition-Based methods to identify this kind patterns. In dataset, if r represents the radius and θ indicates X-axis angle of $\theta = \left(\frac{-\pi}{2}, 0\right), \left(0, \frac{\pi}{2}\right)$ and coordinates (X, Y) is equal to $\begin{cases} x = r.Cos(\theta) \\ y = r.Sin(\theta) \end{cases}$, Datasets $X_1^{\to}$ and $X_2^{\to}$ have been shown in Figure 1.

The Table 1 indicates the coefficients of PSO and GA:

Number Function Evaluation (NFE): Number Function Evaluation (NFE) is a much better measure than CPU performance time to understand the operating speed of the algorithm,because many algorithms introduced recently such as quantum algorithms have operators which are not defined in many common CPUs and applying them needs calculation;So this high runtime of this kind of algorithm cannot be assigned to execution complexity and slowness of the algorithm. NFE is defined as follows:

NFE = number of the original population + (Offspring {number of children + the number of mutants}) * number of iterations;

MATLAB programming environment R2013b version has been used for comparing algorithms. The results of clustering and the cost function based on the number

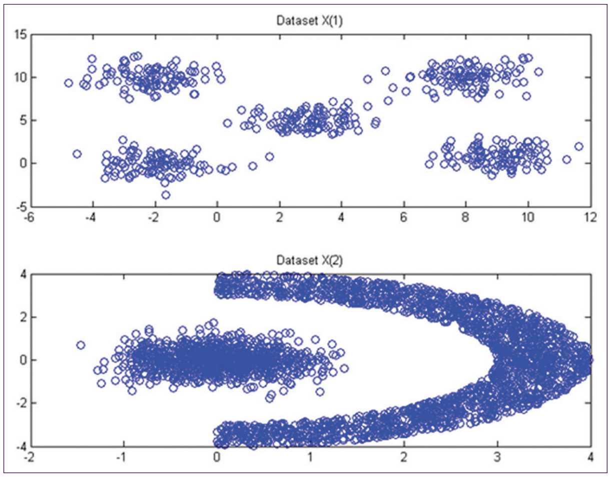


Figure 1: Structure of datasets $X_1^{\to}$ and $X_2^{\to}$

of run cycles on algorithms, K-means optimization based on GA/PSO and the proposed algorithm K-means optimization based on HPG by CS have been shown respectively in Figures 2, 3, 4 and Table 2 for $X_1^{\rightarrow}$ dataset.

Table 1: Values of PSO and GA parameters

PSO parameters

Parameter	Value
Population	100
Max-sub-iteration	2
Inertia weight (ω)	ξ
Personal learning coefficient(C_1)	$\xi\phi_1$
Global Learning coefficient(C_2)	$\xi\phi_2$
damp ($\omega = \omega * damp$)	0.99
Velocity max	0.1*((mix($X^{\rightarrow}$), max($X^{\rightarrow}$))
Velocity min	-max($X^{\rightarrow}$)
Clustering validity indexes	CS

GA parameters

Parameter	Value
Population	100
Max-sub-iteration	1
Cross over percentage	0.8
Mutation percentage	0.3
γ	0.05
Velocity max	0.1*((mix($X^{\rightarrow}$), max($X^{\rightarrow}$))
Velocity min	-max($X^{\rightarrow}$)
Selection pressure(β)	8
Clustering validity indexes	CS

Figure 4 shows the data assignment to each cluster based on the criteria Confusion Matrix by HPG algorithm.

Cluster Cohesion and Separation: Cluster Separation is a standard which indicates the similarity or proximity (in terms of distance criteria) between cluster members and the aim of thecost function is to reduce this amount; on the other hand, separation represents the discrimination or the lowest similarity between different clusters.

In the CS cost function formula (7), the numerator is cohesion criteria and denominator calculates the separation criteria, the following table indicates the average of the two measures and the CS cost function value for Automatic Clustering GA/PSO and algorithms HPG algorithm:

NFE measure in Table 2 has been calculated while algorithm reaches a level of stability (for example, 30 or 40 runs); However, the results show that the proposed HPG algorithm has acted much better than other optimization methods on the $X_1^{\rightarrow}$ dataset and achieved very good results.Figures 2 and 3 and Table 2 show the results of the execution and comparison of algorithms K-means optimization based on GA/PSO and the proposed algorithm K-means optimization based on HPG by CS.

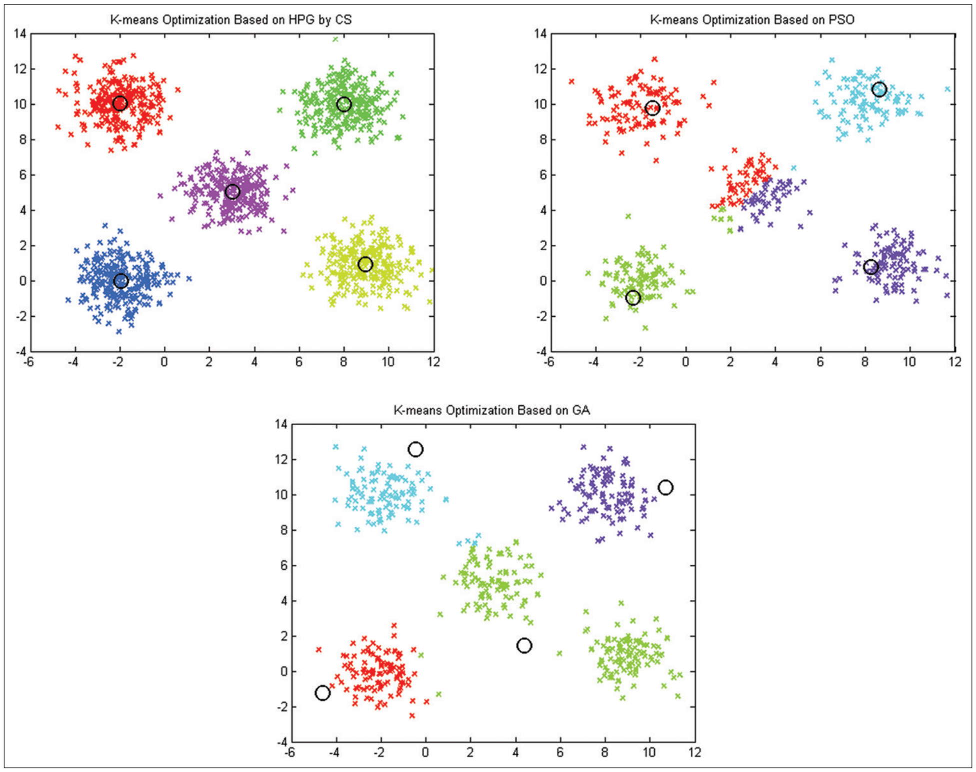


Figure 2: Clustering of $X_1^{\rightarrow}$ dataset respectively by algorithms HPG, PSO and GA

Table 2: The final output of the algorithm (mean of 40 independent run times) in terms of CS cost function and the average NFE

Algorithm	Cohesion	Separation	Number of cluster found	Mean number of NEF require	CS measure
K-means Opt. base on HPG	3.0685	8.3512	5	24010	0.3674
K-means Opt. base on PSO	5.1646	9.4773	4	26600	0.5449
K-means Opt. base on GA	4.0898	7.1102	3	24600	0.5752

Table 3: The final output of the algorithm (mean of 40 independent run times) in terms of CS and the average NFE)

Algorithm	Cohesion	Separation	Number of cluster found	Mean number of NEF require	CS measure
K-means Opt. base on HPG	1.53	2.51	3	10520	0.6036
K-means Opt. base on PSO	1.087	1.94	4	10400	0.5577
K-means Opt. base on GA	1.22	2.18	4	23960	0.563

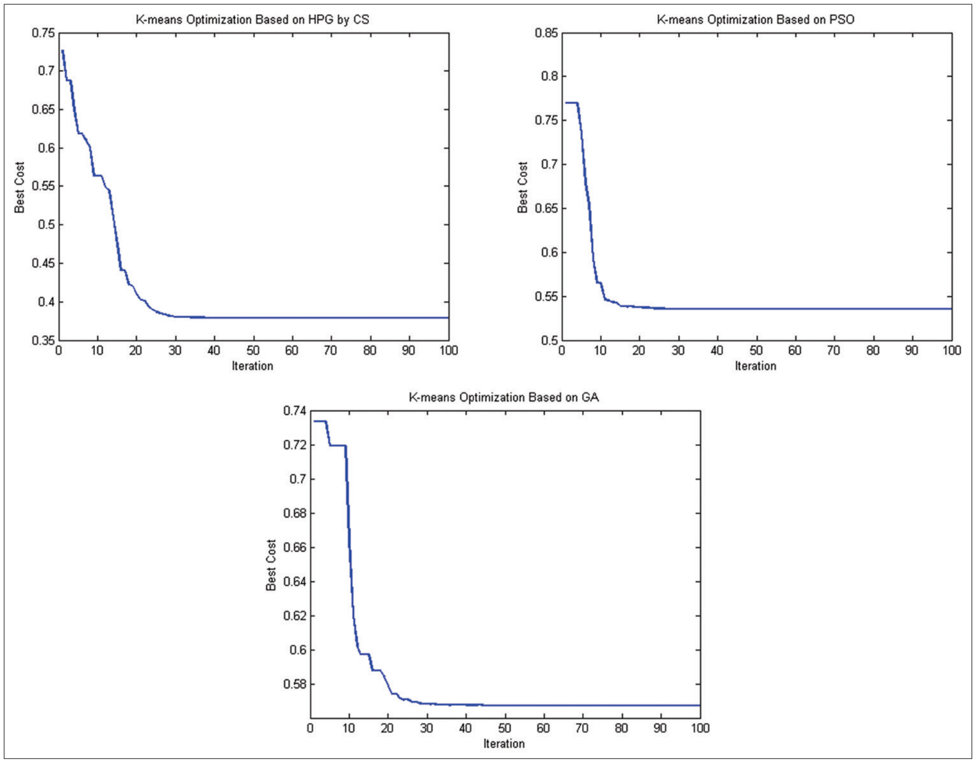


Figure 3: The CS cost function value per 100 cycles for algorithms HPG, PSO, and GA

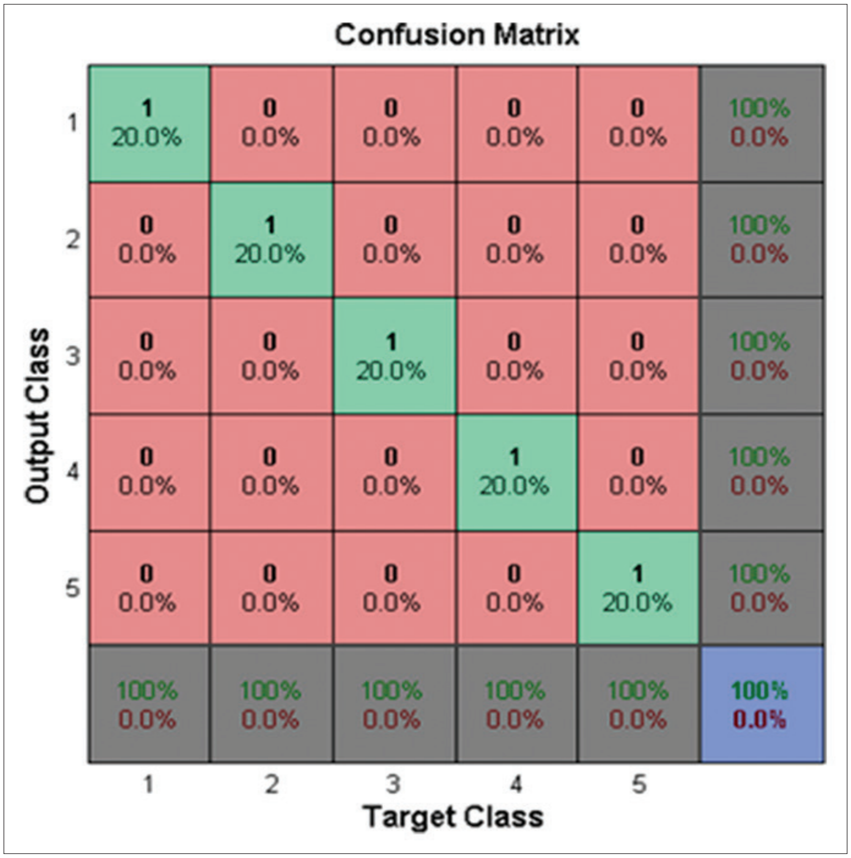
Figures 5 and 6 indicate the attitude to mere optimization and try to minimize the cost function in automatic over-innovative evolutionary algorithms. HPG algorithm, unlike two previous methods, does not only try to optimize and reduce the cost function but is also stopped by finding the best clustering.

Figure 6 shows the data assignment to each cluster based on confusion matrix by HPG algorithm.

Table 3 compares algorithms K-means optimization based on GA/PSO and the proposed algorithm K-means optimization based on HPG by CS in terms of the criteria Cohesion, Separation, CS, and NFE (Figure 7).

CONCLUSIONS

In this paper, optimization of K-means based on HPG algorithm was presented, the basic idea of the proposed approach is the combination of evolutionary and swarm intelligence algorithms to identify the primary centers of clusters and use the CS clustering validity index to activate the cluster centers. In HPG algorithm, at first, the optimal



Confusion Matrix

Output Class \ Target Class	1	2	3	4	5	
1	1 20.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	1 20.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	1 20.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	1 20.0%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 20.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Figure 4. Confusion Matrix calculation for clustering output of HPG algorithm on $X_1^{\rightarrow}$ dataset

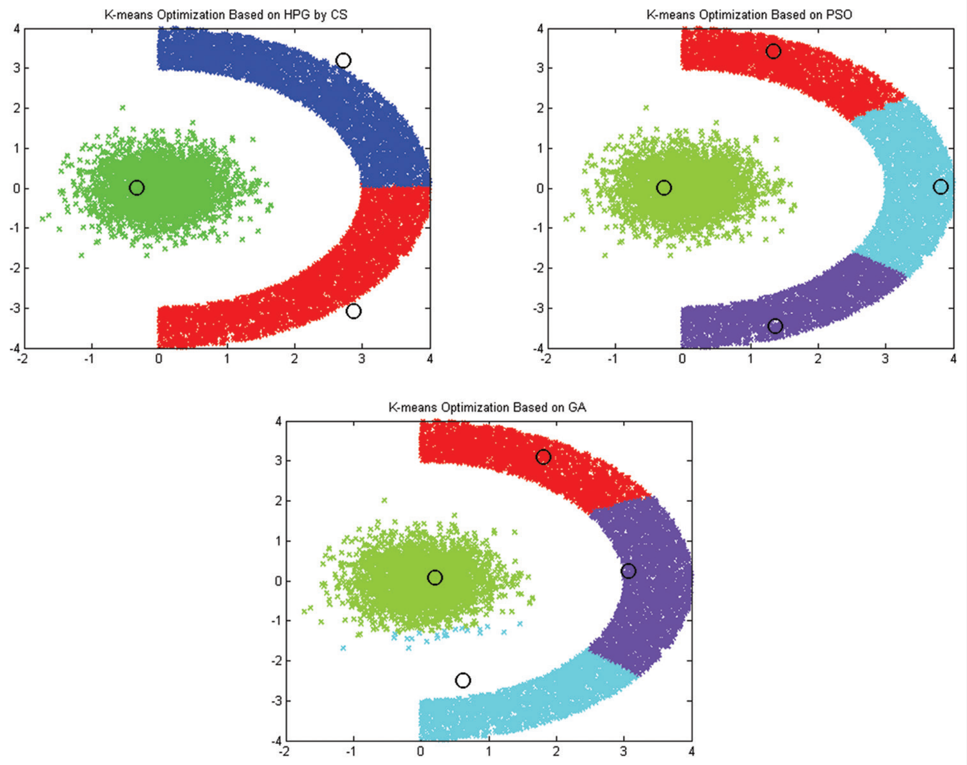


Figure 5. Clustering of $X_2^{\rightarrow}$ dataset respectively by HPG, PSO and GA algorithms

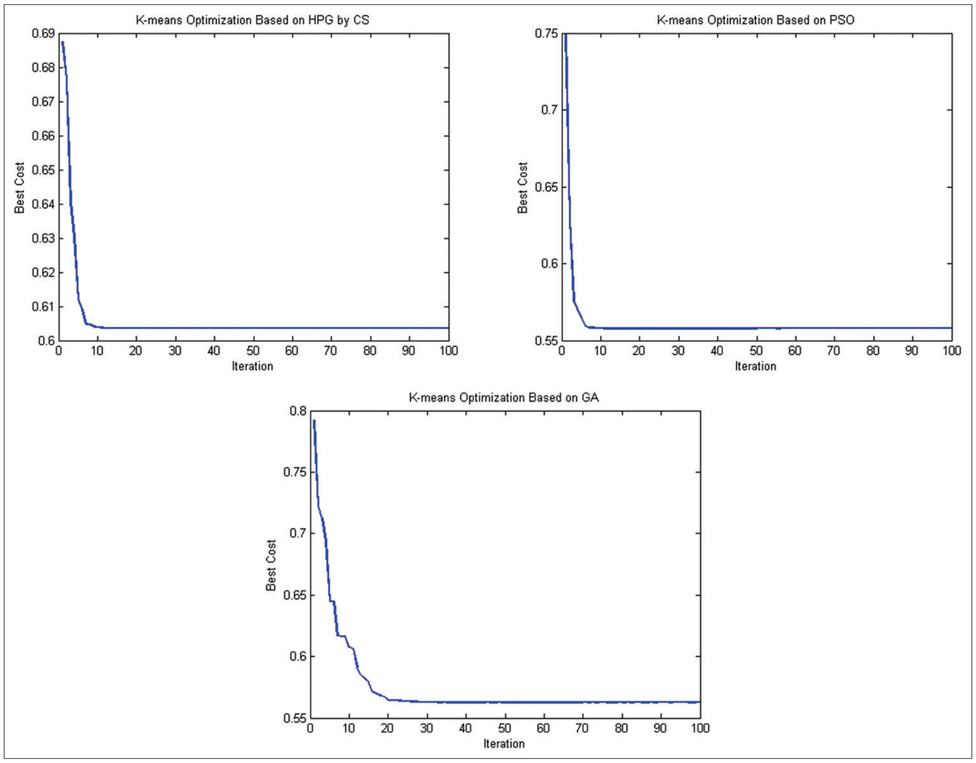


Figure 6: The amount of CS cost function per 100 cycles of algorithms HPG, PSO and GA

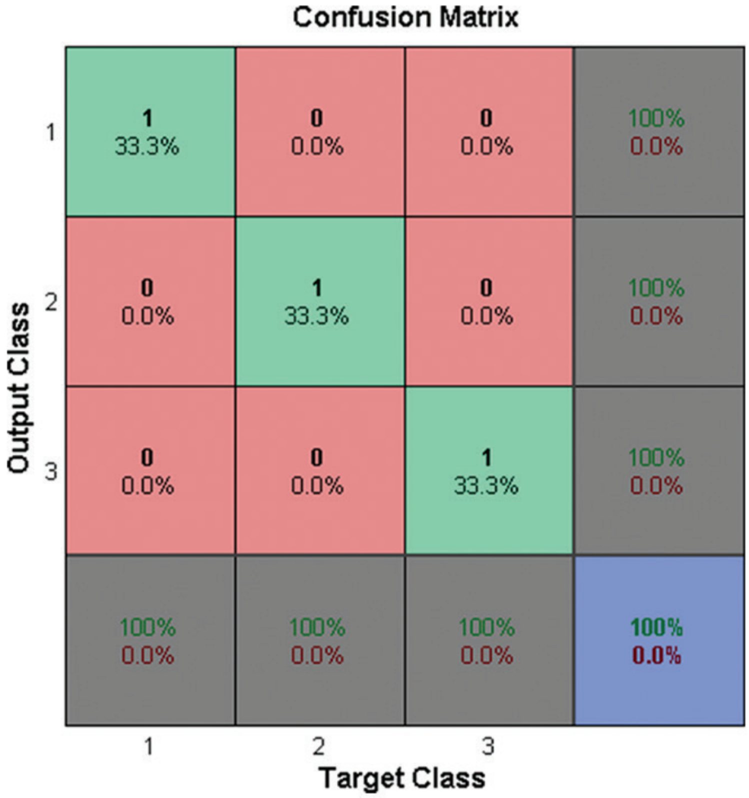


Figure 7: Calculation of confusion matrix for HPG clustering output on $X_2^{\rightarrow}$ dataset

centers of clustering are found based on CS cost function output for each particle population effectively and, finally the most optimal solution with the positions of cluster centers are sent to K-means, this issue solves the problem of random start and fast convergence toward the local optimum of K-means algorithm and other Partition-based algorithms. The combination of PSO algorithm and GA has caused that, HPG algorithm makes use of optimization and connection between the particles together, in addition, by applying the velocity parameter on the population created by GA algorithm, search range and the final answer of this algorithm are improved.

Theoretical results of linear and nonlinear crisp datasets show that HPG algorithm has been converged at an appropriate velocity to the global optimum and in contrast to optimization algorithm K-means based on PSO or GA has not been trapped in alocal optimum. Moreover,comparison of HPG with these methods showed high accuracy in finding the centers of clusters, much more optimal CS cost function and faster implementation of the algorithm according to standard NFE. On the other hand, the traditional K-means algorithm combined with HPG clustering is able to conduct the clustering of complex, large and non-linear datasets with high precision and speed.

In future work, the main purpose will be the extension of HPG algorithm to achieve a comprehensive framework of Data Clustering;Hence, HPG will be used by adding pre-processing steps for outlier detection and reduction of dimensions on thehigh-dimensional data set.In addition, more accurate clustering validity index or combination ofclustering validity indexwill be used as the cost function.

REFERENCES

1. Lin, Y., et al., *K-means optimization clustering algorithm based on particle swarm optimization and multiclass merging*, in *Advances in Computer Science and Information Engineering*. 2012, Springer. p. 569-578.
2. Krishna, K. and M.N. Murty, *Genetic K-means algorithm.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 1999. 29(3): p. 433-439.
3. Lu, Y., et al. *FGKA: A fast genetic k-means clustering algorithm*. in *Proceedings of the 2004 ACM symposium on Applied computing*. 2004. ACM.
4. Lu, Y., et al., *Incremental genetic K-means algorithm and its application in gene expression data analysis.* BMC bioinformatics, 2004. 5(1): p. 1.
5. Das, S., A. Abraham, and A. Konar, *Automatic clustering using an improved differential evolution algorithm.* IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 2008. 38(1): p. 218-237.
6. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. 28(1): p. 100-108.
7. Shi, Y. *Particle swarm optimization: developments, applications and resources*. in *evolutionary computation, 2001. Proceedings of the 2001 Congress on*. 2001. IEEE.
8. Chatterjee, A. and P. Siarry, *Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimization.* Computers & Operations Research, 2006. 33(3): p. 859-871.
9. Whitley, D., *A genetic algorithm tutorial.* Statistics and computing, 1994. 4(2): p. 65-85.
10. Chou, C.-H., M.-C. Su, and E. Lai, *A new cluster validity measure and its application to image compression.* Pattern Analysis and Applications, 2004. 7(2): p. 205-220.