Expert Systems With Applications 79 (2017) 282–295

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Hybrid chemical reaction based metaheuristic with fuzzy c-means algorithm for optimal cluster analysis

Janmenjoy Nayak[a,*], Bighnaraj Naik[b], Himansu Sekhar Behera[c], Ajith Abraham[d,e]

[a] Department of Computer Sc. Engg., Modern Engineering & Management Studies, Balasore-756056,Odisha, India
[b] Department of Computer Application, Veer Surendra Sai University of Technology, Burla, Sambalpur-768018, Odisha, India
[c] Department of Computer Sc. Engg. & Information Technology, Veer Surendra Sai University of Technology, Burla, Sambalpur-768018, Odisha, India
[d] Machine Intelligence Research Labs (MIR Labs), Washington, USA
[e] IT4Innovations - Center of excellence, VSB -Technical University of Ostrava, Czech Republic

ARTICLE INFO

Article history:
Received 13 August 2016
Revised 23 February 2017
Accepted 24 February 2017
Available online 27 February 2017

Keywords:
FCM
Chemical reaction based optimization
K-means
PSO
IPSO
TLBO

ABSTRACT

Hybridization of two or more algorithms has always been a keen interest of research due to the quality of improvement in searching capability. Taking the positive insights of both the algorithms, the developed hybrid algorithm tries to minimize the substantial limitations. Clustering is an unsupervised learning method, which groups the data according to their similar or dissimilar properties. Fuzzy c-means (FCM) is one of the popularly used clustering algorithms and performs better as compared to other clustering techniques such as k-means. However, FCM possesses certain limitations such as premature trapping at local minima and high sensitivity to the cluster center initialization. Taking these issues into consideration, this research proposes a novel hybrid approach of FCM with a recently developed chemical based metaheuristic for obtaining optimal cluster centers. The performance of the proposed approach is compared in terms of cluster fitness values, inter-cluster distance and intra-cluster distance with other evolutionary and swarm optimization based approaches. A rigorous experimentation is simulated and experimental result reveals that the proposed hybrid approach is performing better as compared to other approaches.

1. Introduction

The original evolvement of the term metaheuristic is quite interesting and intends to solve different wide range of problems through heuristic methods. The algorithmic framework of metaheuristic approaches is quite simple as general algorithms, which helps to apply them for solving the real life problems with a few modifications. The broad classification of various optimization algorithms can be of evolutionary based, swarm based, physical based, chemical based, population based and nature based. Although these approaches are quite successful with their own principles, still a single metaheuristic may not be necessarily able to completely obtain the true aspects of both exploration and exploitation. In last decade, a number of advance hybrid methods by combining two or more algorithms have been developed for solving some complex real life problems (Al-Mohair, Saleh, & Suandi, 2015; Amjady, 2007; Carvalho & Freitas, 2004; Chen, Tai, Harrison, & Pan, 2005; Esnaf & Küçükdeniz, 2009; Izakian & Abraham, 2011; Jiang, Li, Yi, Wang, & Hu, 2011; Li, Nguyen, Chen, & Truong, 2015; Sarikprueck, Lee, Kulvanitchaiyanunt, Chen, & Rosenberger, 2015; Shin, Yun, Kim, & Park, 2000; Silva Filho et. al., 2015; Song et. al., 2015; Taşdemir, Milenov, & Tapsall, 2012; Wong & Leung, 2004; Xia et. al., 2015; Yang et. al., 2009; Yang, Huang, & Rao, 2008). The main advantage of using the hybrid methods is to explore the strengths of each of the individual algorithms or procedures for some synergetic performances in combination to both the algorithms (Ting, Yang, Cheng, & Huang, 2015). In such cases, if one will be limited to exploration capabilities, then the other may lead towards exploitation and the outcome of such metaheuristic is quite promising one. Moreover, hybrid approaches result more efficiently in terms of high accuracy or good computational speed.

The earlier days of hybrid metaheuristic algorithms (basically evolutionary and swarm based approaches) such as the development of GA, PSO were interesting in terms of their successful applications. For an instance, the PSO algorithm may be used to optimize the mutation rate of GA. However, evolutionary approaches possess some limitations such as slow convergence, early conver-

* Corresponding author.
E-mail addresses: mailforjnayak@gmail.com, mailtobnaik@gmail.com, mailtohsbehera@gmail.com (J. Nayak), ajith.abraham@ieee.org (A. Abraham).

http://dx.doi.org/10.1016/j.eswa.2017.02.037
0957-4174/

gence, late adaption to the problem etc. As a solution to such limitations, enormous interest has been put forwarded towards some other metaheuristic based optimization approaches such as nature inspired, population based, physical based and chemical based algorithms. The above issues are partially related with the solution diversity, which is produced by the algorithms during the searching process. In case of evolutionary based algorithms, the diversity is sustained by the quality and quantity of organisms at a certain place and time. During the initial stage of searching process, the diversity remains high and while leading towards the global optimum solutions, the diversity may decrease. Although high diversity creates more chances to obtain optimal solution with good accuracy, but is responsible for slower convergence rate. Thus, it is very important to maintain the tradeoff between accuracy with convergence. Moreover, low diversity may result faster convergence and may not guarantee to produce optimal solutions as well as higher accuracy. So, it may be inferred that high diversity leads to exploration and low diversity may not necessarily lead to exploitation. Exploration or diversification is the method of finding the diverse solutions in a search space and exploitation or intensification leads for searching for solution within a local neighborhood of the best solutions i.e. exploitation of the information discovered so far (Fister, Yang, & Brest, 2013). Hence, it is significant to maintain good convergence by cleverly maintaining the exploitation at a correct time and correct place. In addition to this, suitable diversity should be maintained for escaping and jumping out of local optimum positions during the search process. To achieve this, hybridization has been evolved as a key strategy for promoting the diversity and to obtain the global optimum.

Chemical reaction optimization (CRO) (Lam & Li, 2010) is a recent metaheuristic based on the process of some elementary chemical reactions in a chemical system. The advantage of CRO is that, it is free from intricate operators and critical parameter setting like other algorithms and loosely couples chemical reactions with optimization. Also, it uses chemical energy as the heuristic guide to model and optimize the processes. Moreover, it has the advantage of not using any local search methods for refining the search and possesses both local and global searching abilities. Since its inception, this algorithm has proven to be quite successful in solving diversified problems (Alatas, 2011; Asanambigai & Sasikala, 2016; Bechikh, Chaabani, & Said, 2015; Duan & Gan, 2015; Duan & Gan, 2015a; Dutta, Roy, & Nandi, 2016; Lam, Li, & Xu, 2013; Li & Pan, 2012; Li et al., 2015; Naik, Nayak, & Behera, 2016; Truong, Li, & Xu, 2013) in the real world domain. The performance of CRO has outperformed some of the popular evolutionary algorithms. In this work, CRO is used for solving the clustering problem of data mining and its performance is compared with other evolutionary as well as swarm based approaches.

Since last two decades, K-means and FCM are the two most popular techniques for data clustering. K-means is a hard clustering algorithm and is known for its simplicity. The algorithm starts with 'k' number of input parameters and for 'n' number of objects, there will be a possibility of 'k' cluster partitions. The outcome of k-means has more chances for intra-cluster similarity and low chances for inter-cluster similarity. The corresponding object's mean value is treated as the cluster's centroid from which the distance will be measured. In an advance to the working of algorithm, the value of k is to be selected and the performance of the algorithm is purely dependent on the initial centroid for which, chances of getting stuck at local optima is more. On the other side, FCM is quite popular for its use of objective function and is considered as a complete technique. It has been successfully used in various real world applications (Aghabozorgi & The, 2014; Nayak, Naik, & Behera, 2015) such as scheduling, data mining, image processing, feature extraction, nonlinear mappings, engineering domains etc. The reason behind considering the fuzzy based clustering is to avoid the problems of getting stuck at local optima and to find out a solution for choosing optimal cluster center with greater fitness values. However, although normal FCM algorithm exhibits some advantages such as balancing the individual number of cluster points, drifting of small cluster centers to large neighboring cluster centers, presence of fuzzy factor etc., it has also some limitations such as consuming long time for computation, highly sensitive to initial solution, trapping at local minima and sensitive to noisy solutions in case of outliers (Silva Filho, Pimentel, Souza, & Oliveira, 2015).

Additionally, the limitations of some evolutionary based approaches in solving clustering problems such as trapping at local optima, inappropriate cluster center after huge number of iterations, tuning of some specific parameters, slow convergence, producing non optimal solutions in some synthetic datasets etc. have remained open challenges for the present research. In addition to that, during the distance calculation between the object and its nearer cluster center, the result may lead to circular or spherical clusters, which is again very difficult to solve. By considering the above issues, a hybrid metaheuristic based approach (CRO-FCM) is proposed in this work to obtain global optimal solutions and effective clustering. The rest of sections are organized as follows: Section 2 describes the details about various preliminary concepts such as FCM, CRO. The proposed hybrid approach with its working is explained in Section 3. Section 4 outlines the details of parameter settings, result analysis. The results of statistical analysis are mentioned in Sections 5 and 6 concludes the work with future directions.

2. Background study

2.1. Fuzzy c-means algorithm

The FCM algorithm makes use of fuzzy membership function which is used to assign a degree of membership for each class. FCM is able to the form new clusters having close membership values to existing classes of the data points (Dunn, 1974). The technique of FCM relies on three basic operators such as fuzzy membership function, partition matrix and the objective function (Bezdek, 2013). FCM is used to partition a set of 'N' clusters through minimization of the objective function (Zadeh, 1965) with respect to the fuzzy partition matrix.

$$J(U,V) = \sum_{i=1}^{C}\sum_{j=1}^{N} {u_{ij}}^{m} \left\| x_j - v_i \right\|^2 \tag{1}$$

where 'x_j' denotes the jth cluster point, 'v_i' represents the ith cluster center, $u_{i,j}$ is the membership value of 'x_j' w.r.t. cluster 'i', 'm' denotes the fuzzy controlling parameter i.e. for the value '1', it will tend to hard partition and for the value of '∞' it tends towards the complete fuzziness and $\|\|$ represents the norm function.

The iterative method is used to compute the membership function (Lin, Huang, Kuo, & Lai, 2014) and cluster center as:

$$U_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\left\| x_j - v_i \right\|}{\left\| x_j - v_k \right\|} \right)^{\frac{2}{m-1}} \right]^{-1} \tag{2}$$

$$v_i = \sum_{j=1}^{N} {u_{ij}}^{m} x_j \Big/ \sum_{j=1}^{N} {u_{ij}}^{m} \text{ where } i \geq 1, i \leq c \tag{3}$$

The working procedure of FCM algorithm is illustrated in Fig. 1.

2.2. Chemical reaction optimization

The potential solutions for CRO are simulated in terms of a molecule which has kinetic energy, potential energy with some

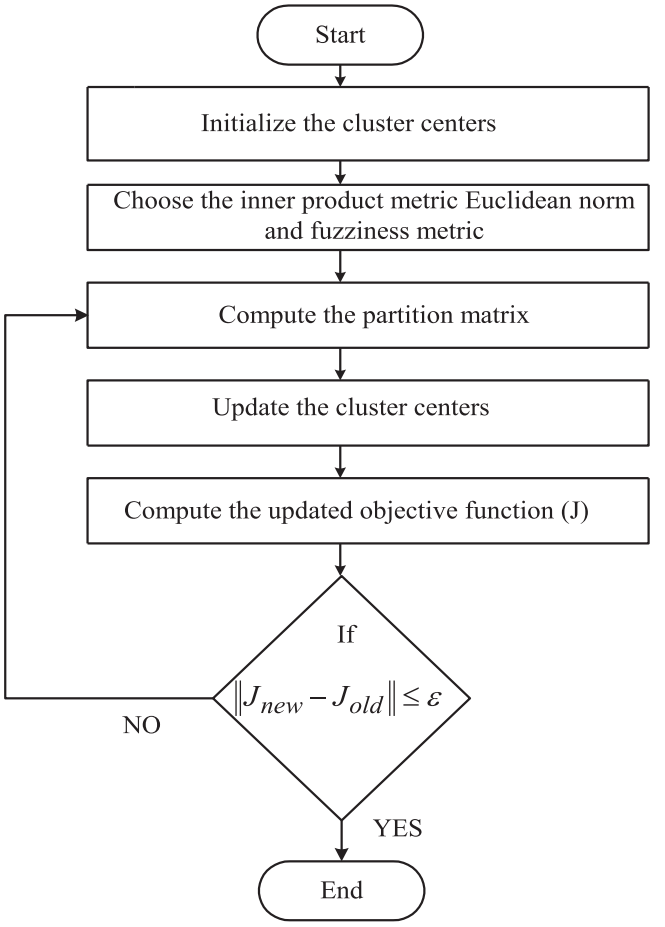

Fig. 1. Steps of FCM algorithm.

other characteristics. The working procedure of CRO depends on four elementary chemical reactions such as intermolecular collision, synthesis, on wall collision and decomposition. In a chemical reaction system, the potential energy leads towards the minimum values at some balanced state. This behavior of potential energy is treated as the value of the objective function and the potential energy is the fitness value for the required process.

The transformation of one or more chemical compounds into another chemical compounds with different physical/chemical properties by means of simultaneous breaking of and making of chemical bonds is defined as chemical reaction (Lam et al., 2013). In every chemical reaction, the formation of product/products is accompanied with the release of energy to the surroundings, whereas; in order to allow the chemical compounds to react with each other some short energy should supplied initially from the surroundings to reach the active state of the chemical reaction. Depending on the amount of energy supplied for breaking of bonds within the reactant molecules and the amount of energy released during formation of products, the heat of enthalpy of product formation is divided into two categories, endothermic and exothermic. Exothermic process is associated with final heating effect for product formation, whereas; a net cooling effect is observed for endothermic process. As every chemical entity possess a definite amount of internal energy, related to all the bond energies within that chemical entity, reactants have some specific energy as well as product have some specific energy (Nayak, Naik, & Behera, 2015a). If the products are formed with low energy as compared to the reactants, then cooling will be observed, whereas; for reverse case heating will be observed.

2.2.1. Intermolecular collision

It means the collision of two or more reactant molecule with each other resulting in the formation of different product.

Example: 2NO (Nitric Oxide) + $2H_2$ (Hydrogen) → N_2 (Nitrogen) + $2H_2O$ (Water)

In general:

If $w1 + w2 \rightarrow w1' + w2'$ is a reaction and if $KE(w1) + KE(w2) + PE(w1) + PE(w2) < PE(w1') + PE(w2')$, then the new chemical system $(w1' + w2')$ is accepted, else rejected. Here KE and PE are the Kinetic & Potential Energy respectively, w1, w2, w1', w2' are the reactants.

2.2.2. On wall collision

It means collision occurs in between the reactant molecule and wall of a container. It is also known as uni-molecular reaction where the molecule undergoes exothermic process with the removal of some amount of heat energy.

Example: $2O_3$ (Ozone) → $3O_2$(Oxygen)

In general:

If $w \rightarrow w'$ is a reaction and if $KE(w) + PE(w) < PE(w')$, then the new chemical system (w') is accepted, else rejected.

2.2.3. Synthesis

A synthesis reaction occurs when more than one molecules or reactants pooled each other to produce a sole compound.

Example: 2Li (Lithium) + Cl_2 (Chlorine) → 2LiCl (Lithium Chloride)

In general:

If $w1 + w2 \rightarrow w$ is a reaction and if $KE(w1) + KE(w2) + PE(w1) + PE(w2) < PE(w)$, then the new chemical system (w) is accepted, else rejected.

2.2.4. Decomposition

Decomposition reaction occurs when a molecule dissociates into different fragments on collision with the wall of a container. It is the reverse process of synthesis reaction.

Example: NH_4NO_2 (Ammonium Nitrite) → N_2 (Nitrogen) + $2H_2O$ (Water)

In general:

If $w \rightarrow w1' + w2'$ and if $KE(w) + PE(w) < PE(w1') + PE(w2')$, then the new chemical system $(w1' + w2')$ is accepted, else rejected.

The four components along with the detail working of CRO algorithm has been described in Fig. 2.

3. Proposed approach

This section deals with the detail procedures of the proposed hybrid CRO-FCM approach. Solving the problem in optimization point of view, the objective may be for the minimization of overall deviation between the cluster partitions. So, distance is to be calculated among the total distance between the clusters and their relevant cluster center (the objective function value of FCM algorithm). Each candidate solution (w_i) in the population is represented through k number of cluster centers. As per the size of the population (i.e. n), the w_i is generated randomly. Here uniform distribution process has been adopted for the generation of random weight values for the candidate solution (set of weight) in the population. This allows every possible solution for the inclusion in the population with equal chance. The objective of this work is to choose best cluster centers (w_i) from the population to improve intra-cluster and inter-cluster distance. Traditionally, FCM starts with random initialization of cluster centers in which chances of getting good cluster centers is low. To avoid this, in this

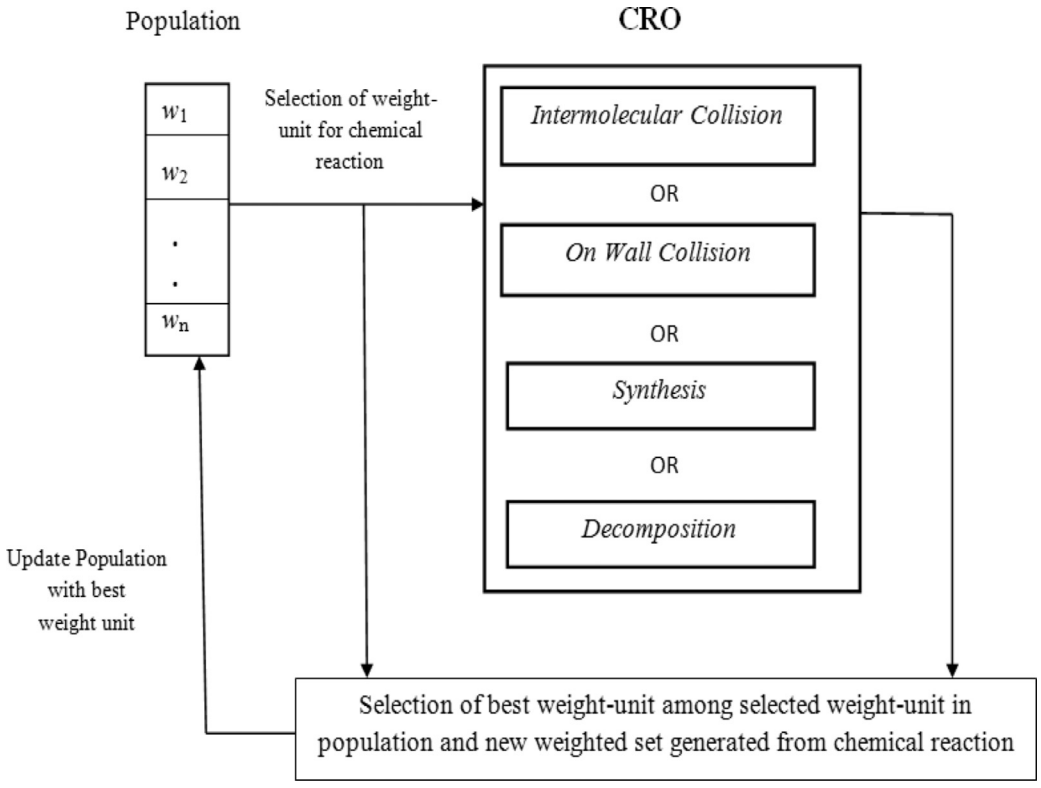

Fig. 2. Working of CRO algorithm for population updation.

work a recently developed chemical reaction based metaheuristic such as CRO is used to select the best cluster centers which may improve the performance of FCM algorithm. Usually, FCM algorithm uses initial predefined number of cluster centers which are to be initialized randomly and gradually, these cluster centers are updated in FCM iterations based on membership values. Here, each cluster center is represented as a vector and each value in this vector are generated randomly using uniform distribution, in order to explore the imaginary solution space uniformly.

Here all the candidate solutions w_1, w_2,...w_n are considered as 'n' number of molecules in a chemical system. CRO basically works on four elementary chemical reactions such as intermolecular collision, synthesis, on wall collision and decomposition. Each of these reactions generates more molecules (w_i) from the candidate solutions and finally stable molecules from the resultant molecules are chosen by eliminating unstable molecules from the population. The complete working model is illustrated in Fig. 3.

The types of reactions to be selected are decided by the randomly generated value of r_1 and r_2. Both r_1 and r_2 will generate a value in between 0 to 1. If $r_1 > 0.5$ and $r_2 > 0.5$, then decomposition reaction is triggered. In decomposition reaction, i^{th} molecule (w_i) is selected and decomposed into two resultant molecules (w_i' and w_i"). Fitness of these molecules (w_i, w_i' and w_i") are computed by using Eq. (4). Best molecule among these is selected based on their fitness values, which replace 'w_i' in the population. If $r_1 > 0.5$ and $r_2 <= 0.5$, then on wall collision reaction takes place. In this reaction, i^{th} molecule (w_i) is selected and bounced back in the form of new molecule (w_i') after collision with wall of the container. Fitness of these two molecules (w_i & w_i') are computed by using Eq. (4). Best molecule among these is selected based on their fitness values, which replace w_i in the population. If $r_1 \leq 0.5$ and $r_2 \leq 0.5$, then intermolecular collision reaction occurs. In this reaction, two number of molecules (w_i & w_j) are selected and produces two more new molecules (w_i' & w_j'). Fitness of these four molecules (w_i, w_j, w_i', w_j') is computed by using Eq. (4). Best two molecules among these are selected based on their fitness values, which replace w_i & w_j in the population. If $r_1 \leq 0.5$ and $r_2 > 0.5$, then synthesis reaction occurs. In this reaction, two numbers of molecules (w_i & w_j) are selected and are synthesized into a new molecule (w_k). Fitness of these three molecules (w_i, w_j, w_k) are computed by using Eq. (4). Best two molecules among these are selected based on their fitness values, which replace w_i & w_j in the population. In these processes, selecting best molecule based on their fitness is analogous to keeping stable molecules in a chemical system. By replacing initial candidate solutions with improved candidate solutions produced from these four reactions, new population is generated. If the stopping criteria are met, then stop the process and return the best candidate solution (cluster centers). Else, the old population is to be updated with new population. These reactions are triggered conditionally and a single reaction is invoked in the CRO iterations. Here, r1 and r2 are two random numbers generated with uniform distribution which allows the triggering of all four elementary reactions with equal probability. This process is iterated until there is no further significant improvement in the solution vectors or the maximum number of iteration is reached.

After having the best cluster centers produced from the above CRO process, these centers are set as initial cluster center to FCM algorithm and then FCM algorithm will be simulated to get best optimal solutions to clustering. The mean of the clusters is calculated along with the fitness of each cluster center using Eq. (4).

$$F(w_i) = \frac{k}{(\sum_{j=1}^{m}\sum_{r=1}^{n}(o_r - c_{i,j})^2) + d} \tag{4}$$

In Eq. (4), 'w_i' indicates the i^{th} candidate solution in the population 'P', $F(w_i)$ represents the fitness of the solution 'w_i', 'n' is the number of instance in the dataset, 'o_r' is the r^{th} instance in the dataset, 'm' is the number of cluster centers in 'w_i', '$c_{i,j}$' is

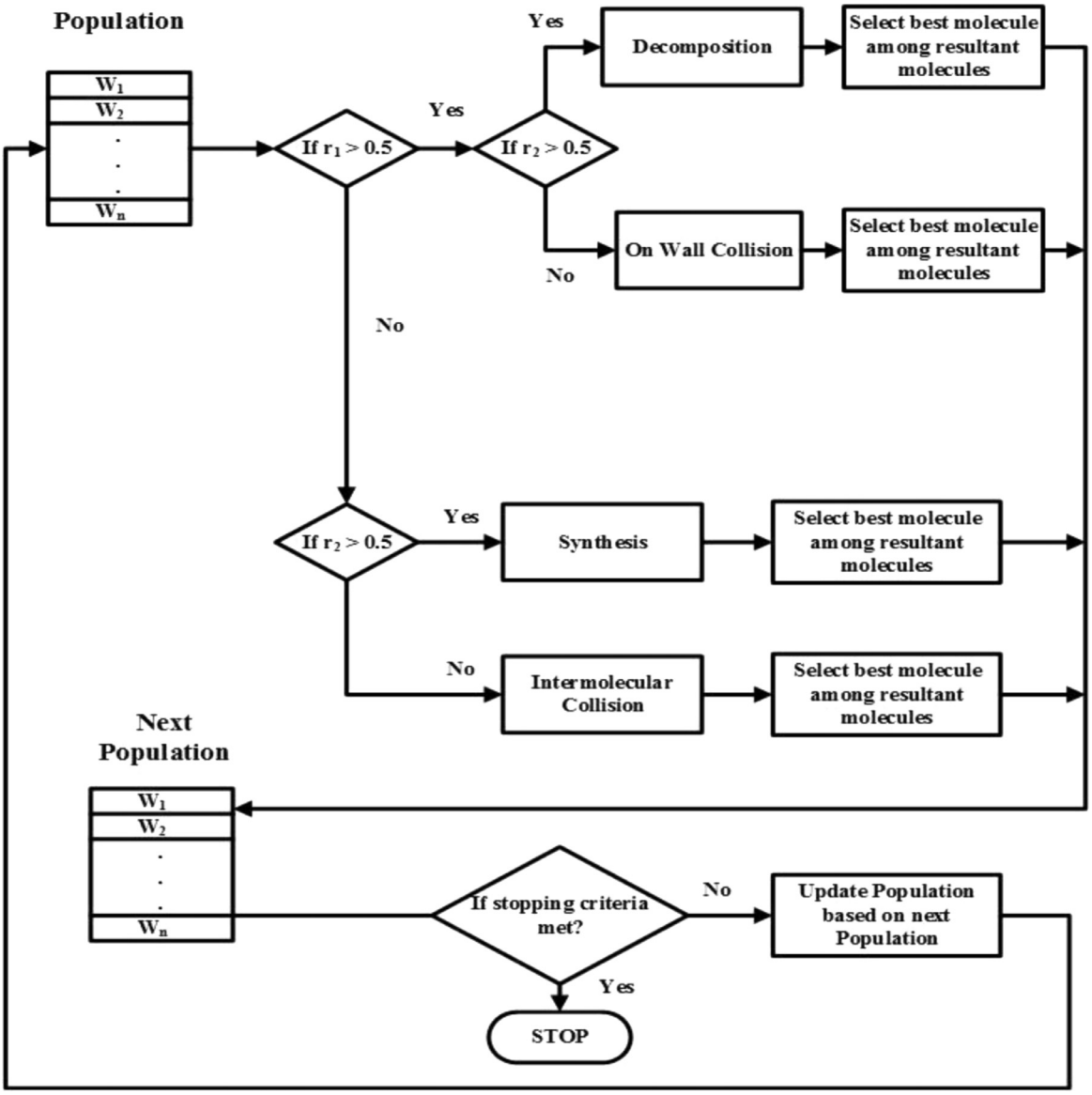

Fig. 3. Working model of CRO-FCM approach.

the j^{th} cluster center of i^{th} (w_i) solution from 'P', '*k*' is a positive constant, '*d*' is a small valued constant (Ahmadyfard & Modares, 2008). In Eq. (4), the denominator part is the sum of cluster distance from cluster centers, which is usually a decimal or nearer to decimal value and in fact, this cluster distance changes from data sets to data sets. Based on considered data sets and obtained fitness values, some other alternative suitable values may be considered for these parameters. In order to have a complete focus on the performance comparison and making the test impartial, these required parameters 'k', 'd' and 'm' (fuzziness factor) have been set to fixed values. However, these values may be suitably changed depending upon problem domain. While changing these values in the allowed range, we have observed homogeneous effects (either increasing/decreasing) on the performance of all the considered models, i. e. if algorithm A is better than B in one parameter setting, then it is remained as it is in other parameter settings.

Here '*F*' is a function to evaluate the generalized solutions called fitness function. 'o_r - $c_{i,j}$' is the Euclidean distance from the object 'o_r' to '$c_{i,j}$' (cluster center). The main objective of using Eq. (4) to minimize the intra-cluster distance. i.e. when the intra-cluster distance is low, the value of the objective function will be high. The detailed pseudo code of the of the proposed CRO-FCM algorithm is described in Algorithm 1. In the algorithm based on the conditions, any of the four reactions such as decomposition (Algorithm 3), on wall collision (Algorithm 4), synthesis (Algorithm 5) and intermolecular collision (Algorithm 6) will be triggered. The best cluster center based on the fitness values will be evaluated through the BestMoleculeSelection procedure (Algorithm 2).

4. Experimental analysis and simulation results

This section deals with the experimental analysis and details about the required parameters to be set for the simulation of the hybrid CRO-FCM approach. The section is divided into following subsections.

4.1. Dataset information

The datasets used for this experiment are listed in Table 1. There are twelve number of real world benchmark datasets and

Algorithm 1: Pseudo-code of the proposed hybrid CRO-FCM algorithm.

1. Initialize the population of Cluster centers $P = \{w_1, w_2, \ldots..w_N\}$ (candidate solutions) randomly, where $w_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,m}\}$, N is the size of the population and m is the number of clusters in the population
2. ***while**(1)*
   ***For** i = 1:1:Lp-1*
   r1 = rand(1);
   r2 = rand(1);
   ***If** (r1 > 0.5)*
   [w] = BestMoleculeSelection (P);
   ***If** (r2 > 0.5)*
   [P] = Decomposition (w,P);
   else
   P = OnWallCollision (w,P);
   end
   else
   ***If** (fitnessmol(i) > fitnessmol(i + 1))*
   w4 = P(i,:);
   mxi1 = i;
   else
   w4 = P(i + 1,:);
   mxi1 = i;
   end
   [w] = BestMoleculeSelection (P);
   ***If** (r2 > 0.5)*
   P = Synthesis (w4,w5,P);
   else
   P = InterMolecularCollision (w4,w5,P);
   end
   end
   end
   iter = iter + 1;
   ***If** (iter == max_iter)*
   break;
   end
   End
3. Compute the fitness of each cluster center w_i in the population by using Eq. (4) and find out the best solution.
4. Rank the cluster centers based on their fitness, obtain the best cluster center.
5. Initialize the FCM center with position of the best cluster center.
6. Then using this center, iterate the FCM algorithm.
   Repeat
   Update the membership matrix by Eq. (2).
   Refine the cluster centers by Eq. (3).
   Do until it meets the convergence criteria
7. **Exit**

Algorithm 2: BestMoleculeSelection Method.

1. ***function** [w] = **BestMoleculeSelection** (P)*
2. *for i = 1:1:nor,where nor is number of row in P*
3. *Calculate fitness of cluster center 'w' i.e. P(i,:) from population P by using Eq. (4).*
4. ***end***
5. *Select cluster center 'best' on the basis of maximum fitness.*
6. *Assign w = best.*
7. ***end***

Algorithm 3: Decomposition method.

1. ***function** [P] = **Decomposition** (w,P)*
2. *w1 = w + (−1 + (1 - −1).*rand(1));*
3. *w2 = w + (−1 + (1 - −1).*rand(1));*
4. *Compute fitness of cluster centers w, w1 & w2.*
5. *Select best two among these to replace two number of cluster centers from P.*
6. ***end***

Algorithm 4: Onwallcollision method.

1. ***function** P = **OnWallCollision** (w,P)*
2. *i = rand(1);*
3. ***if** (i > 0.5)*
4. *w3 = w + (−1 + (1 - −1).*rand(1));*
5. ***else***
6. *w3 = w-(−1 + (1 - −1).*rand(1));*
7. ***end***
8. *Compute fitness of cluster centers w & w3.*
9. *Select best one among these to replace one number of cluster centers from P.*
10. ***end***

Algorithm 5: Synthesis method.

1. ***function** P = **Synthesis** (w4,w5, P)*
2. ***for** k = 1:1:L, where 'L' is length of w4 or w5*
3. *i = rand(1);*
4. ***if** (i > 0.5)*
5. *w6(1,k) = w4(1,k);*
6. ***else***
7. *w6(1,k) = w5(1,k);*
8. ***end***
9. ***end***
10. *Compute fitness of cluster centers w4, w5 & w6.*
11. *Select best two among these to replace two number of cluster centers from P.*
12. ***end***

Algorithm 6: InterMolecularCollision method.

1. ***function** [w7,w8] = **InterMolecularCollision** (w4,w5)*
2. *i = rand(1);*
3. ***if** (i > 0.5)*
4. *w7 = w4 + (−1 + (1 - −1).*rand(1));*
5. *w8 = w5 + (−1 + (1 - −1).*rand(1));*
6. ***else***
7. *w7 = w4 - (−1 + (1 - −1).*rand(1));*
8. *w8 = w5 - (−1 + (1 - −1).*rand(1));*
9. ***end***
10. *Compute fitness of cluster centers w7, w8,w4 & w5.*
11. *Select best two among these to replace two number of cluster centers from P.*
12. ***end***

Table 1
Information about the datasets.

Sl. No	Datasets	No. of pattern	No. of clusters	No. of attributes
1.	Iris	150	3	4
2.	Lenses	24	3	4
3.	Haberman	306	3	3
4.	Balance scale	625	3	4
5.	Wisconsin breast cancer	699	3	10
6.	Contraceptive method choice	1473	3	9
7.	Hayesroth	132	3	5
8.	Robot navigation	5456	4	2
9.	Spect heart	80	2	22
10.	Glass	214	6	9
11.	Wine	178	3	13
12.	Artificial dataset	600	3	2
13.	Lung cancer	32	3	56

Table 2
Comparison on clustering metric results of K-means, FCM, TLBO and CRO algorithms.

	Clustering metric results of K-means, FCM, TLBO & CRO Algorithms			
	K-Means	*FCM*	*TLBO*	*CRO*
Iris	0.012395396	0.012738542	0.012800898	0.03155815
Lenses	0.339904827	0.381339952	0.352281023	0.340239036
Haberman	0.000317745	0.000316547	0.000337213	0.000373982
Balance scale	0.002573387	0.003332606	0.002912077	0.003172302
Wisconsin breast cancer	7.25935E-14	7.48861E-14	6.49028E-14	7.38806E-14
Contraceptive method choice	7.80139E-05	7.69432E-05	7.91398E-05	7.950183E-05
Hayesroth	4.59807E-05	4.43056E-05	4.629092E-05	4.680829E-05
Robot navigation	0.001583094	0.002000381	0.002220018	0.002298267
Spect heart	0.069341756	0.077804472	0.071902265	0.069992846
Glass	0.181666666	0.214233564	0.219003011	0.197523655
Wine	4.83293E-07	4.6507E-07	4.723039E-07	4.782358E-07
Artificial dataset	4.94137E-06	4.91855E-06	4.940607E-06	4.947219E-06
Lung cancer	2.30381921	2.51627612	2.820912221	2.921829028

one artificial dataset has been considered and the details about the dataset such as number of pattern, number of clusters, number of attributes are illustrated in the Table.

4.2. Simulation environment

The developing environment for the proposed method is MATLAB 9.0 on a system with an Intel Core Duo CPU T5800, 2 GHz processor, 2 GB RAM and Microsoft Windows-2007 OS.

4.3. Experimental results and performance analysis

For the experiment of the proposed hybrid approach and to compare it with other approaches twelve real world bench mark datasets from UCI machine learning repository (Bache & Lichman, 2013) and one artificial dataset has been considered. The detail properties of the datasets such as number of clusters, patterns and attributes are explained in Table 1. The performance of the proposed approach is compared with some popular methods such as K-means, FCM, TLBO, CRO and some hybrid methods such as GA-K-means, PSO-K-means, TLBO-K-means, ETLBO-K-means, GA-FCM, PSO-FCM. The performance of the proposed hybrid approach is assessed and compared using the criteria such as cluster fitness (cluster metric), intra-cluster distance, inter-cluster distance, error rate and number of iterations. The rate of error (Hatamlou, 2013) is calculated by dividing the number of misplaced objects with the total number of objects belongs to one particular dataset (Eq. (5)).

$$\text{Rate of error} = \text{no. of misplaced objects} / \text{total no of objects in the dataset} \quad (5)$$

For the implementations of FCM algorithm, the fuzzy controlling parameter or the weighting factor 'm' has been set to 2. The details of implementation procedures and parameters set for the approaches such as K-means, GA-K-means, PSO-K-means, TLBO-K-means and ETLBO-K-means may be found in Kanungo, Nayak, Naik, and Behera (2016), Nayak, Naik, Kanungo, and Behera (2015b) and Nayak, Kanungo, Naik, and Behera (2016). For PSO, the values of the acceleration coefficients c1 and c2 are set to 1.4, inertia weight is set in between 1.8 and 2; for TLBO, and the teaching factor (T_F) is set to 1 or 2 with equal probability. For implementing all the algorithms, 50 number of independent runs is simulated and the results of 1400 number of iterations are indicated in Tables 2–4. The performance comparison among the approaches such as K-means, FCM, TLBO and CRO is shown in Table 2 and the fitness values of four hybrid algorithms based on K-means clustering such as GA-K-means, PSO-K-means, TLBO-K-means and ETLBO-means are represented in Table 3. The comparison of the proposed hybrid CRO-FCM with some other well known approaches such as GA-FCM & PSO-FCM in terms of their clustering metric values is indicated in Table 4. From these three tables (Tables 2–4), it may be inferred that the performance of the proposed hybrid chemical reaction based approach is quite better as compared to all the considered methods in this work. The fitness values of CRO-FCM in all the considered real world benchmark datasets are high (except the artificial dataset) than the hybrid K-means approaches and other hybrid FCM based approaches. In case of artificial dataset, the fitness value of CRO-FCM is 4.98865E-06, PSO-FCM is 4.97987E-06 and GA-FCM is 4.96589E-06. In this case, CRO-FCM is performing better than all the considered approaches, but ETLBO-K-means produces slightly better result than CRO-FCM. Other than this, the proposed hybrid method has higher superiority over the other approaches. The performance of fitness metric by using the methods such as K-means, FCM, GA-FCM, PSO-FCM and CRO-FCM for various datasets has been illustrated in Figs. 4–11.

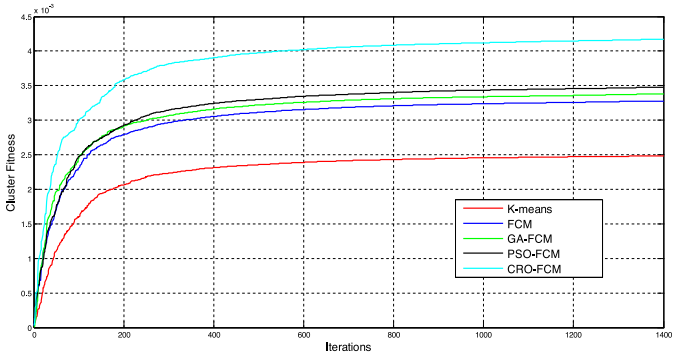

Fig. 4. Comparison of fitness value of CRO-FCM with other methods on Balance dataset.

Table 5 represents the mean and standard deviations of the intra-cluster and inter-cluster distances for all the considered methods along with the proposed method by considering the datasets such as iris, Haberman, breast cancer, Hayesroth, glass and wine datasets respectively. The obtained mean values for intra-cluster distance are 2.2803 for iris dataset, 27.498 for Haberman dataset, 12.2487 for Wisconsin breast cancer dataset, 3.1022 for Hayesroth dataset, 62.8638 for glass dataset, and 442.7833 for wine dataset respectively. In all these considered cases, the proposed hybrid approach is quite efficient and produced less intra-cluster distance than the other approaches.

Further, for better comparison among the methods, the average of mean values of all the methods of both inter and intra cluster distances for all the considered data sets are calculated. Both the results of inter and intra cluster distances are ranked according to

Table 3
Comparison on clustering metric results of GA-K-means, PSO-K-means, TLBO-K-means and ETLBO-K-means algorithms.

	Clustering metric results of various hybrid K-means Algorithms			
	GA-K-means	*PSO-K-means*	*TLBO-K-means*	*ETLBO-K-means*
Iris	0.013826351	0.014528017	0.014635644	0.014724565
Lenses	0.351735427	0.360239542	0.443532685	0.444011111
Haberman	0.000328364	0.000348162	0.000388888	0.000384213
Balance scale	0.002628475	0.002810827	0.003725464	0.003722008
Wisconsin breast cancer	7.26287E-14	7.28928E-14	7.54648E-14	7.54844E-14
Contraceptive method choice	8.03819E-05	8.20198E-05	8.25254E-05	8.25291E-05
Hayesroth	4.70825E-05	4.73918E-05	4.75140E-05	4.77111E-05
Robot navigation	0.001828362	0.001898018	0.002746987	0.002858946
Spect heart	0.072648917	0.076041565	0.084362513	0.084384626
Glass	0.182496522	0.191000011	0.265555551	0.263018566
Wine	4.84222E-07	4.85339E-07	4.88326E-07	4.88416E-07
Artificial dataset	4.95447E-06	4.96647E-06	4.98822E-06	4.98888E-06
Lung cancer	2.670915679	2.76985463	2.966975354	2.985648755

Table 4
Comparison on clustering metric results of GA-FCM, PSO-FCM and CRO-FCM algorithms.

Datasets	*Clustering metric results of various hybrid FCM Algorithms*		
	GA-FCM	*PSO-FCM*	*CRO-FCM*
Iris	0.014154986	0.014624876	0.014859984
Lenses	0.390354824	0.425698354	0.495692544
Haberman	0.000330542	0.000372865	0.000442243
Balance scale	0.003425487	0.003535478	0.004238763
Wisconsin breast cancer	7.50236E-14	7.52487E-14	7.59489E-14
Contraceptive method choice	8.13254E-05	8.20398E-05	8.84358E-05
Hayesroth	4.71657E-05	4.74493E-05	4.89622E-05
Robot navigation	0.002258745	0.002454781	0.003446324
Spect heart	0.079365885	0.080456544	0.086687254
Glass	0.235687998	0.248023652	0.479652264
Wine	4.85985E-07	4.86258E-07	4.91633E-07
Artificial dataset	4.96589E-06	4.97987E-06	4.98865E-06
Lung cancer	2.729946254	2.863599423	3.269744214

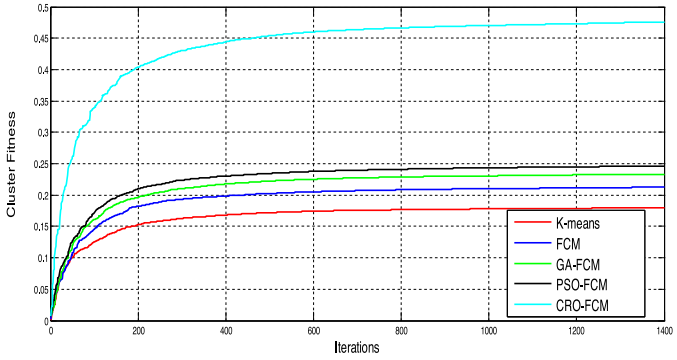

Fig. 5. Comparison of fitness value of CRO-FCM with other methods on Glass dataset.

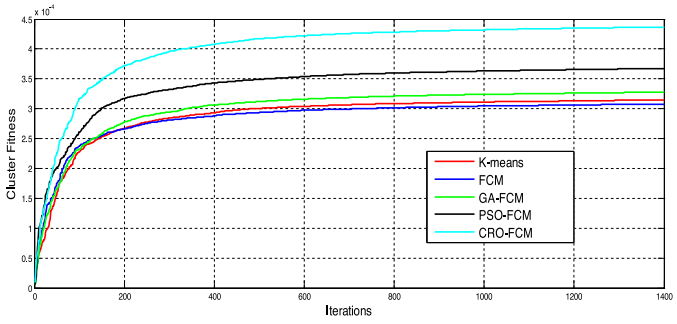

Fig. 6. Comparison of fitness value of CRO-FCM with other methods on Haberman dataset.

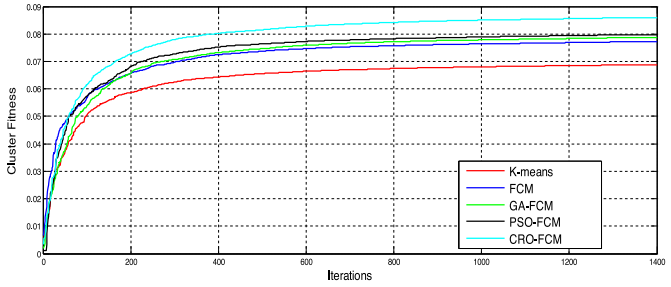

Fig. 7. Comparison of fitness value of CRO-FCM with other methods on Spect heart dataset.

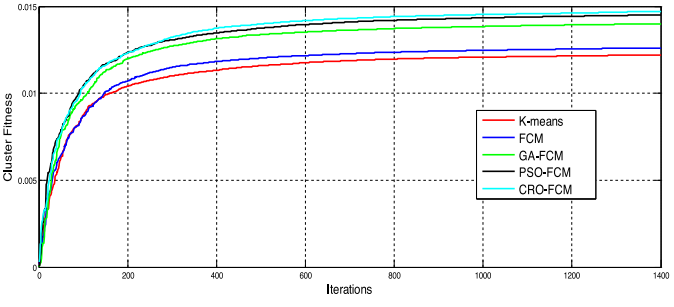

Fig. 8. Comparison of fitness value of CRO-FCM with other methods on Iris dataset.

Table 5
Comparison of Intra-cluster and Inter-cluster distances among all the Clustering algorithms on thirteen datasets.

Dataset			*Inter and Intra cluster values of all the methods*										
			K-means	*FCM*	*TLBO*	*CRO*	*GA-K-means*	*PSO-K-means*	*TLBO-K-means*	*ETLBO-K-means*	*GA-FCM*	*PSO-FCM*	*CRO-FCM*
Iris	**Inter**	**Mean**	1.5982	1.6355	1.7994	1.7924	1.6724	1.6837	1.7246	1.7284	1.9327	1.7303	1.7826
		Std. Dev	0.4187	0.4239	0.0044	4.66E-16	0.4463	0.4868	0.483	0.4852	0.4239	0.4106	0.3689
	Intra	**Mean**	2.5936	2.5224	2.5644	2.5076	2.4826	2.4264	2.4032	2.4006	2.3621	2.3517	2.2803
		Std. Dev	0.1812	0.1781	0.0447	4.66E-16	0.1552	0.1504	0.142	0.1411	0.1524	0.1683	0.1464
Lenses	**Inter**	**Mean**	422.81	438.82	467.22	469.39	434.83	442.74	478.9	482.48	532.85	544.28	583.72
		Std. Dev	11.2	18.03	21.35	24.83	26.46	25.93	35.83	36.22	46.66	48.53	54.28
	Intra	**Mean**	512.62	476.27	378.92	356.99	320.88	311.3	356.99	355.21	307.64	318.72	302.11
		Std. Dev	23.88	12.99	11.28	11.68	22.87	19.36	12.31	12.26	10.62	14.82	10.29
Haberman	**Inter**	**Mean**	13.16	14.296	12.6179	13.3972	12.3972	14.2614	16.2913	16.31	18.348	19.63	19.9
		Std. Dev	1.2416	1.5838	0	1.4758	1.1682	1.646	1.6894	1.6896	1.7493	1.8453	1.8824
	Intra	**Mean**	41.84	41.92	44.6816	42.3893	39.3328	38.692	36.2942	36.29	35.268	32.238	27.498
		Std. Dev	3.2622	3.2646	2.18E-14	3.2421	3.2421	3.2238	3.2019	3.2042	3.1942	3.1256	2.7608
Balance scale	**Inter**	**Mean**	1027.86	1243.49	1228.83	1282.23	1296.93	1383.21	1392.88	1397.8	1458.4	1472.3	1483.47
		Std. Dev	435.99	572.62	528.2	538.22	520.82	578.26	583.71	592.64	629.72	677.24	672.02
	Intra	**Mean**	1478.34	1340.43	1380.52	1322.93	1305.27	1309.29	1302.37	1302.15	1293.4	1177.27	1128.29
		Std. Dev	522.93	478.38	492.39	438.29	446.7	428.69	418.92	412.62	406.2	400.84	357.2
Wisconsin breast cancer	**Inter**	**Mean**	3.3658	3.379	2.9359	4.3893	4.5692	3.5683	4.7624	4.7687	4.962	4.98	5.5763
		Std. Dev	2.0269	2.0489	1.9658	1.8269	1.8453	1.3822	1.8896	1.8897	1.9222	1.9421	2.3896
	Intra	**Mean**	17.1748	16.2534	19.7297	16.6149	16.5	16.384	15.07	15.04	14.872	14.2453	12.2487
		Std. Dev	0.4176	0.4012	0.5276	0.3938	0.3808	0.3682	0.3402	0.34	0.2873	0.265	0.1985
Contraceptive method choice	**Inter**	**Mean**	2414.3	2786.97	2902.442	3109.22	3566.2	3878.59	3939.3	3983.8633	3922.2	3962.52	3982.34
		Std. Dev	6.6212	18.2901	16.4432	21.4432	24.8624	26.3277	23.4423	23.4955	28.333	29.0443	29.8373
	Intra	**Mean**	4872.209	4209.84	4277.214	4199.492	4086.02	3977.49	3399.22	3372.8832	3103	3006.87	2994.2
		Std. Dev	2.3992	4.232	16.3	5.3372	1.7832	1.0322	6.4322	6.1139	2.3436	2.038	1.5826
Hayesroth	**Inter**	**Mean**	1.6531	1.6382	1.4124	1.6838	1.7184	1.768	1.9208	1.9204	2.4892	2.6931	2.8631
		Std. Dev	0.2634	0.247	0.2153	0.2863	0.2966	0.3257	0.4698	0.4693	1.8763	1.9164	1.9642
	Intra	**Mean**	3.3782	3.3987	3.5823	3.3539	3.3486	3.3208	3.29	3.2745	3.16	3.1243	3.1022
		Std. Dev	1.293	1.3091	1.3492	1.2752	1.2501	1.2208	1.1932	1.1917	1.1248	1.1196	1.1073
Robot navigation	**Inter**	**Mean**	488.72	568.82	562.82	574.91	583.45	588.28	593.37	598.73	632.8	674.39	683.29
		Std. Dev	32.28	56.77	54.22	57.27	54.38	56.61	59.52	59.92	67.82	75.28	86.38
	Intra	**Mean**	387.43	328.72	347.62	319.28	312.44	304.92	304.24	304.18	299.9	234.8	214.66
		Std. Dev	48.62	34.6	42.91	31.18	31.63	27.82	26.82	26.55	22.35	23.73	17.2
Spect heart	**Inter**	**Mean**	2319.92	2361.98	2382.9	2446.72	2677.82	2683.39	2739.62	2402.28	2389.8	2429.82	2557.83
		Std. Dev	1283.49	1288.2	1312.94	1452.22	1563.38	1620.33	1677.3	1682.43	1762.4	1782.34	1783.93
	Intra	**Mean**	1088.29	1023.61	1062.81	1007.38	883.27	837.62	828.27	818.76	862.38	819.62	802.28
		Std. Dev	437.82	432.81	446.82	407.27	342.77	322.72	304.72	303.17	318.29	302.9	300.72
Glass	**Inter**	**Mean**	70.8539	70.2936	70.1062	70.38	70.976	74.6432	76.5793	76.82	77.489	78.4359	79.4923
		Std. Dev	0.962	0.2983	0.4931	0.6934	1.1638	2.3976	3.4682	3.4765	2.6347	3.593	3.76348
	Intra	**Mean**	70.37821	70.2234	70.2934	70.31896	70.121	69.5324	66.46	66.1	64.694	63.9868	62.8638
		Std. Dev	0.63892	0.58315	0.37314	0.6052	0.2	0.19634	0.162284	0.1602	0.2111	0.14936	0.11632
Wine	**Inter**	**Mean**	293.3158	290.693	265.4835	270.3789	294.786	296.635	298.5238	298.6281	306.8	309.748	318.779
		Std. Dev	47.1146	46.4937	31.4793	5.82E-14	48.1796	48.8934	49.1394	49.2653	58.248	39.2782	62.2463
	Intra	**Mean**	456.7149	457.249	459.1782	462.338	454.492	451.492	456.9425	455.8342	451.49	458.796	442.783
		Std. Dev	2.8698	2.882	2.8938	2.9683	2.7692	2.7692	2.682	2.6652	2.22	2.1483	2.1128
Artificial dataset	**Inter**	**Mean**	13.3097	14.392	13.3029	14.8862	18.2399	18.8862	19.8724	19.7231	27.309	28.3902	34.5874
		Std. Dev	2.6721	2.0489	2.9342	3.832	4.8402	6.3231	6.7922	5.4937	5.3387	5.3099	6.332
	Intra	**Mean**	28.4898	24.9821	26.34	23.4032	23.087	22.33	21.2891	21.0429	18.201	17.3056	14.2215
		Std. Dev	3.921	3.7381	3.5342	3.3021	3.3231	1.3238	2.3776	2.1092	1.2302	2.2342	1.1309
Lung cancer	**Inter**	**Mean**	3.1861	3.2281	3.3038	3.3694	3.8621	3.884	3.8904	3.8942	4.21	5.8201	5.8849
		Std. Dev	0.3822	0.3405	0.1243	0.0287	0.3251	0.4602	0.5721	0.5801	0.5997	0.6802	0.725
	Intra	**Mean**	7.5107	7.4263	7.5542	7.4722	7.4201	7.4094	7.5308	7.5296	7.0711	6.2844	4.9287
		Std. Dev	0.2954	0.2741	0.2994	0.1242	0.1224	0.1205	0.1186	0.1173	0.1028	0.1127	0.1092
Mean average of all inter cluster distances			544.1579	599.972	608.8595	635.5959	689.804	722.426	735.97192	714.534323	721.5	733.441	**750.732**
Rank of Mean average of all inter cluster distances			11	10	9	8	7	4	2	6	5	3	**1**
Mean average of all intra cluster distances			689.7669	615.604	621.616	602.6515	578.82	565.554	523.10537	520.053462	497.19	473.509	**462.421**
Rank of Mean average of all intra cluster distances			11	9	10	8	7	6	5	4	3	2	**1**

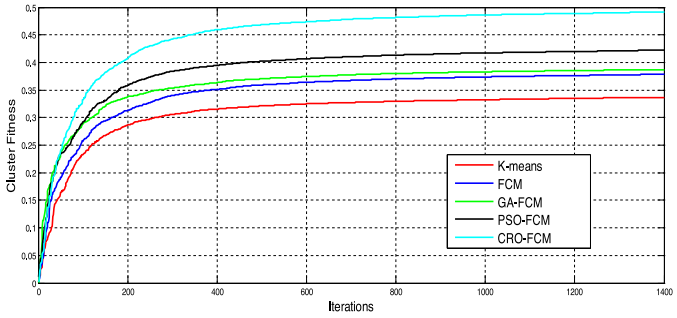

Fig. 9. Comparison of fitness value of CRO-FCM with other methods on Lense dataset.

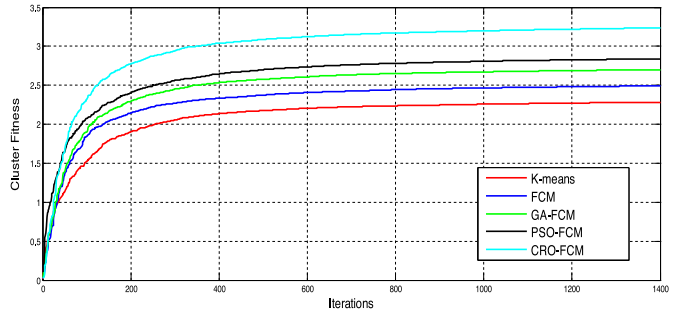

Fig. 10. Comparison of fitness value of CRO-FCM with other methods on Lung cancer dataset.

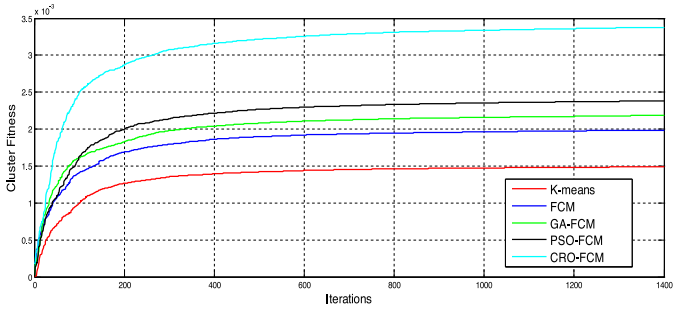

Fig. 11. Comparison of fitness value of CRO-FCM with other methods on robot navigation dataset.

their performance on mean values and revealed that the proposed method is superior as compared to others (Fig. 12). Also, the performance of the proposed work is compared with some other reported literatures (Baral & Behera, 2013; Satapathy & Naik, 2014) and found that, the mean values of both intra-cluster as well as inter-cluster distance is better as compared to them. The reason behind the consideration of mean values is that the considered algorithms such as GA, PSO are stochastic algorithms, in which the value of each iteration may vary. Also, as the variation on the simulation results in each iteration is directly dependent on the number of iterative looping for a particular algorithm so, some effective performance measures such as fitness value and average number of iterations required to obtain the optimal solution has been considered in this work. The average number of iterations for resulting optimal solution is demonstrated in Table 6 and it is found that the proposed method is able to get the optimal solution in less number of iterations than others. The rate of error for all the considered algorithms is calculated by using Eq. (5) and is listed in Table 7. From Table 7, it is clear that the average error rate of the proposed method is quite less than other approaches. In both the Tables 6 & 7, the last row indicates the ranking of various algorithms based on their average number of iterations and average number of rate of errors over all the considered datasets respectively. For both the considered metrics i.e. average number of iterations and average number of rate of errors, it is found that the rank of proposed method is '1', which clearly divulges it's superiority over the other methods.

5. Statistical significance

In this section, to show the significance differences among all the methods, statistical tests such as Friedman test, Iman-Davenport test, Holm test are performed by considering all the cases. First, Friedman test as well as Iman-Davenport test is employed to test the significance differences among the results in various clustering techniques. Then, if significant differences are found, post hoc test (Holm test) is carried out by considering one group as control group against all the other algorithms. These three tests are carried out by considering two results such as fitness values and error rate of all the considered algorithms. The value of α is set as 0.01 which is the confidence level in all the considered cases. In Friedman test, the ranks are assigned to all the algorithms and the best performing algorithm is assigned as rank 1. In this work, the proposed CRO-FCM is ranked as 1 and the rest of the algorithms are ranked as per their performance. Accordingly, the average ranks for all the considered cases are calculated and those ranks are used to compute the Friedman statistics. The detailed description about all these tests is presented in Nayak et al. (2015a).

Fitness values of seven methods presented in Tables 3 & 4 has been considered for assigning the ranks based on their performance and accordingly the Friedman test is carried out on them. The reason behind considering seven methods and not considering the rest four methods such as K-means, FCM, TLBO & CRO is that, these four are not able to produce at least a nearer range of values to the proposed method after 50 generations, for which results of the proposed method is far better than them. Table 8 represents the ranks assigned to seven methods against all the thirteen datasets based on the fitness metric and the average ranks are calculated accordingly. In Table 9, the statistical values of Friedman's test and Iman-Davenport test are compared with the critical value, as a result all the null hypothesis are rejected. Table 10 reports the results of Holm test, where the proposed CRO-FCM is considered as the control algorithm which is to be tested with all the other algorithms. The results reported in the table divulge that, CRO-FCM performs quite better in maximum cases, except last two where the marginal variability is very less. Out of six, in four cases the null hypothesis is rejected. So, based on the fitness metric, the proposed method is statistically significant over other methods. The density plot having (6, 72) degree of freedom based on the fitness metric is shown in Fig. 13.

Similar tests are performed on all the seven methods based on their rate of error. The proposed CRO-FCM is ranked as 1 and then successively ranked all other algorithms based on their performance and the average ranks are computed in Table 11. Based on the average ranks, both Friedman's statistic and Iman-Davenport statistical values are calculated and compared with the critical value. For both the cases the null hypothesis is rejected based on their rate of error (Table 12).

During the Holm test, based on the z values, corresponding p values are computed for all the algorithms and the results are reported in Table 13. From the obtained results, the null hypothesis is rejected in four out of six, which demonstrates better statistical significance of the proposed method than others based on the error rates. The density plot having (6,72) degree of freedom based on the rate of error is shown in Fig. 14.

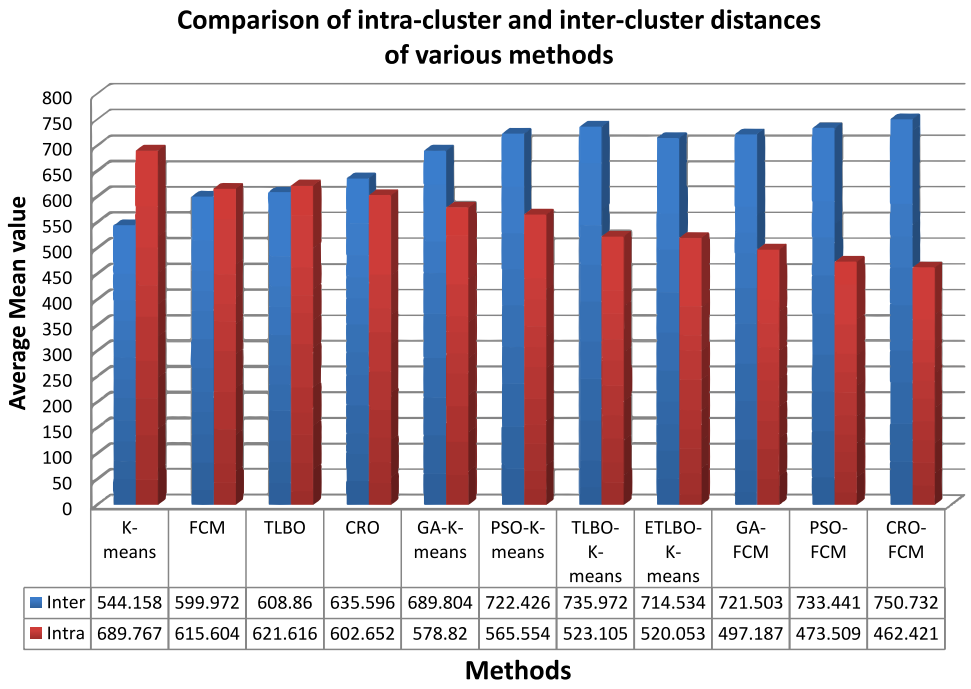

Fig. 12. Comparison of inter and intra cluster distance of various methods.

Table 6
Comparison among all the considered clustering approaches based on average number of iterations to obtain optimal solutions.

Dataset	*Average no. of iterations required by various Clustering methods*										
	K-means	*FCM*	*TLBO*	*CRO*	*GA-K-means*	*PSO-K-means*	*TLBO-K-means*	*ETLBO-K-means*	*GA-FCM*	*PSO-FCM*	*CRO-FCM*
Iris	73	34	78	27	33.4	28.65	26.2	26	27.4	26.6	14.3
Lenses	48.76	44	49.4	38.9	42.44	34.96	34.7	32.4	37.66	30.2	28.4
Haberman	106.22	92.83	71.3	23.73	67.9	62.3	72.30	71	79.5	49.2	49.66
Balance scale	227.8	186.2	142.4	113.2	156.93	112.93	106	106	132.4	128.6	88.2
Wisconsin breast cancer	137.4	70.28	91.23	44.8	110.8	114.72	99.9	99	87.6	84.39	62.3
Contraceptive method choice	227.8	187.6	134.2	147.2	122.2	128.6	110.9	110.9	124.5	102.6	84
Hayesroth	48.4	56.3	40.7	40.9	39.4	38.82	30.11	30	43.2	36.9	29.3
Robot navigation	112.70	130.8	109.2	89.65	97.30	92	86.49	83.20	72.2	78.4	70.29
Spect heart	167.82	120.2	138	128.3	118.29	123.8	120.4	114.8	98	92.4	88.6
Glass	84.2	42	65.9	73.2	54.6	56.8	50.8	49.81	40.2	40	32.80
Wine	142.3	64.9	118.20	42.29	94.2	86.9	76.40	78.9	56.9	52.4	52
Artificial dataset	237.22	186.4	78.30	92	184.34	162.6	124.99	124	132.9	99.2	102.5
Lung cancer	36.9	14	67.8	30.6	34.2	26.66	44.2	42.9	12.84	18.8	9.52
Ranking of Algorithms on Average number of iterations over all the dataset	***11***	***10***	***9***	***3***	***8***	***7***	***5***	***4***	***6***	***2***	***1***

Table 7
Comparison among all the considered clustering approaches based on average rate of error.

Dataset	*Error rate of various Clustering methods in (%)*										
	K-means	*FCM*	*TLBO*	*CRO*	*GA-K-means*	*PSO-K-means*	*TLBO-K-means*	*ETLBO-K-means*	*GA-FCM*	*PSO-FCM*	*CRO-FCM*
Iris	13.34	12.43	14.3	13.28	12.3	12.20	11.87	11.8	11.20	11.1	9.18
Lenses	20.16	20.12	20.10	19.39	20.9	18.03	17.20	17.1	19.8	18.34	16.40
Haberman	9.29	9.10	8.2	8.18	9.16	9.05	8.12	8.09	7.62	7.35	7.19
Balance scale	34.20	31.4	32.10	31.3	30.90	30.7	30.4	30.14	27.62	27.18	27.10
Wisconsin breast cancer	7.39	7.2	7.27	7.08	7.15	7.02	7.01	7	7.28	6.94	6.91
Contraceptive method choice	53.47	53.12	54.17	52.89	52.76	52.68	52.38	52.36	52.32	52.30	52.21
Hayesroth	13.46	13.42	13.38	13.35	13.28	13.24	13.21	13.22	13.26	13.20	13.16
Robot navigation	23.4	23.37	23.31	23.25	23.38	23.34	23.29	23.28	23.46	23.25	23.22
Spect heart	6.28	6.23	6.48	6.21	6.22	6.21	6.20	6.14	6.18	6.17	6.13
Glass	38.39	38.26	42.36	38.11	38.37	38.32	38.34	38.27	38.35	38.31	38.14
Wine	31.11	31.26	31.24	31.08	31.09	31.06	31.07	31.08	30.21	30.18	30.12
Artificial dataset	24.28	24.16	23.19	24.10	23.15	23.13	23.12	23.10	24.08	23.28	23.02
Lung cancer	18.52	18.24	18.16	18.15	18.50	18.46	18.20	18.16	18.49	18.11	18.06
Ranking of Algorithms on Average rate of error over all the dataset	***10***	***9***	***11***	***7***	***8***	***6***	***5***	***3***	***4***	***2***	***1***

Table 8
Assignment of Friedman's rank to all the considered algorithms based on fitness metric.

Sl. No. of dataset	*Assigned ranks to Clustering metric results of various Clustering algorithms*						
	GA-K-means	*PSO-K-means*	*TLBO-K-means*	*ETLBO-K-means*	*GA-FCM*	*PSO-FCM*	*CRO-FCM*
1.	0.013826351(7)	0.014528017(5)	0.014635644(3)	0.014724565(2)	0.014154986(6)	0.014624876(4)	0.014859984(1)
2.	0.351735427(7)	0.360239542(6)	0.443532685(3)	0.444011111(2)	0.390354824(5)	0.425698354(4)	0.495692544(1)
3.	0.000328364(7)	0.000348162(5)	0.000388888(2)	0.000384213(3)	0.000330542(6)	0.000372865(4)	0.000442243(1)
4.	0.002628475(7)	0.002810827(6)	0.003725464(2)	0.003722008(3)	0.003425487(5)	0.003535478(4)	0.004238763(1)
5.	7.26287E-14(7)	7.28928E-14(6)	7.54648E-14(2)	7.54844E-14(3)	7.50236E-14(5)	7.52487E-14(4)	7.59489E-14(1)
6.	8.03819E-05(7)	8.20198E-05(5)	8.25254E-05(3)	8.25291E-05(2)	8.13254E-05(6)	8.20398E-05(4)	8.84358E-05(1)
7.	4.70825E-05(7)	4.73918E-05(5)	4.75140E-05(3)	4.77111E-05(2)	4.71657E-05(6)	4.74493E-05(4)	4.89622E-05(1)
8.	0.001828362(7)	0.001898018(6)	0.002746987(3)	0.002858946(2)	0.002258745(5)	0.002454781(4)	0.003446324(1)
9.	0.072648917(7)	0.076041565(6)	0.084362513(3)	0.084384626(2)	0.079365885(5)	0.080456544(4)	0.086687254(1)
10.	0.182496522(7)	0.191000011(6)	0.265555551(2)	0.263018566(3)	0.235687998(5)	0.248023652(4)	0.479652264(1)
11.	4.84222E-07(7)	4.85339E-07(6)	4.88326E-07(3)	4.88416E-07(2)	4.85985E-07(5)	4.86258E-07(4)	4.91633E-07(1)
12.	4.95447E-06(7)	4.96647E-06(5)	4.98822E-06(3)	4.98888E-06(1)	4.96589E-06(6)	4.97987E-06(4)	4.98865E-06(2)
13.	2.670915679(7)	2.76985463(5)	2.966975354(3)	2.985648755(2)	2.729946254(6)	2.863599423(4)	3.269744214(1)
Average Ranks	**7**	**5.53**	**2.69**	**2.23**	**5.46**	**4**	**1.07**

Table 9
Results of Friedman and Iman-Davenport Test based on the fitness metric.

Test	*Statistical Value*	*Obtained Critical Value*	*Hypothesis*
Friedman	74.30	3.06	Rejected
Iman-Davenport	240.97	3.06	Rejected

Table 10
Results of Holm Test based on the fitness metric (Control algorithm: CRO-FCM).

i	*Method*	*z-value*	*p-value*	$\alpha/(m-i)$	*Hypothesis*
6	***GA-K-means***	6.26	0	0.01	*Rejected*
5	***PSO-K-means***	5.27	6.814477E-8	0.005	*Rejected*
4	***GA-FCM***	5.18	1.109419E-7	0.003	*Rejected*
3	***PSO-FCM***	3.46	0.00027	0.0025	*Rejected*
2	***TLBO-K-means***	1.91	0.028067	0.002	*Not Rejected*
1	***ETLBO-K-means***	1.37	0.085343	0.0016	*Not Rejected*

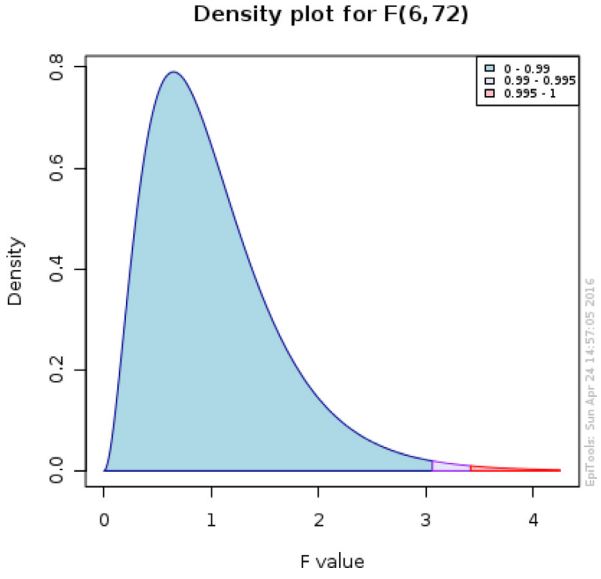

Fig. 13. Density Plot on the Degree of Freedom (6,72) based on the Fitness metric.

6. Conclusion and future directions

Applying hybrid metaheuristic algorithms for solving complex problems as well as modeling and simulation of these algorithms to achieve optimal solutions has been a recent interest among the researchers. In this work, a hybrid parameter free chemical reaction based fuzzy-c-means clustering algorithm is proposed to obtain optimal cluster centers. The experiments reported in the paper considered twelve real world benchmark datasets and one artificial dataset. The experimental results of the proposed method are compared with some hybrid approaches such as GA-K-means, PSO-K-means, TLBO-K-means, ETLBO-K-means, GA-FCM, PSO-FCM and other benchmark models such as K-means, FCM, TLBO, CRO. The simulation results have shown that the hybrid CRO-FCM proposed in this paper achieved better results regarding various performance criteria such as fitness metric, intra-cluster distance, inter-cluster

Table 11
Assignment of Friedman's rank to all the considered algorithms based on rate of error.

Sl. No. of Dataset	*Assigned ranks to error rates of various Clustering algorithms*						
	GA-K-means	*PSO-K-means*	*TLBO-K-means*	*ETLBO-K-means*	*GA-FCM*	*PSO-FCM*	*CRO-FCM*
1.	12.3 (7)	12.20 (6)	11.87 (5)	11.8 (4)	11.20 (3)	11.1 (2)	9.18 (1)
2.	20.9 (7)	18.03 (4)	17.20 (3)	17.1 (2)	19.8 (6)	18.34 (5)	16.40 (1)
3.	9.16 (7)	9.05 (6)	8.12 (5)	8.09 (4)	7.62 (3)	7.35 (2)	7.19 (1)
4.	30.90 (7)	30.7 (6)	30.4 (5)	30.14 (4)	27.62 (3)	27.18 (2)	27.10 (1)
5.	7.15 (6)	7.02 (5)	7.01 (4)	7 (3)	7.28 (7)	6.94 (2)	6.91 (1)
6.	52.76 (7)	52.68 (6)	52.38 (5)	52.36 (4)	52.32 (3)	52.30 (2)	52.21 (1)
7.	13.28 (7)	13.24 (5)	13.21 (3)	13.22 (4)	13.26 (6)	13.20 (2)	13.16 (1)
8.	23.38 (6)	23.34 (5)	23.29 (4)	23.28 (3)	23.46 (7)	23.25 (2)	23.22 (1)
9.	6.22 (7)	6.21 (6)	6.20 (5)	6.14 (2)	6.18 (4)	6.17 (3)	6.13 (1)
10.	38.37 (7)	38.32 (4)	38.34 (5)	38.27 (2)	38.35 (6)	38.31 (3)	38.14 (1)
11.	31.09 (7)	31.06 (4)	31.07 (5)	31.08 (6)	30.21 (3)	30.18 (2)	30.12 (1)
12.	23.15 (5)	23.13 (4)	23.12 (3)	23.10 (2)	24.08 (7)	23.28 (6)	23.02 (1)
13.	18.50 (7)	18.46 (5)	18.20 (4)	18.16 (3)	18.49 (6)	18.11 (2)	18.06 (1)
Average Ranks	**6.69**	**5.07**	**4.30**	**3.30**	**4.92**	**2.69**	**1**

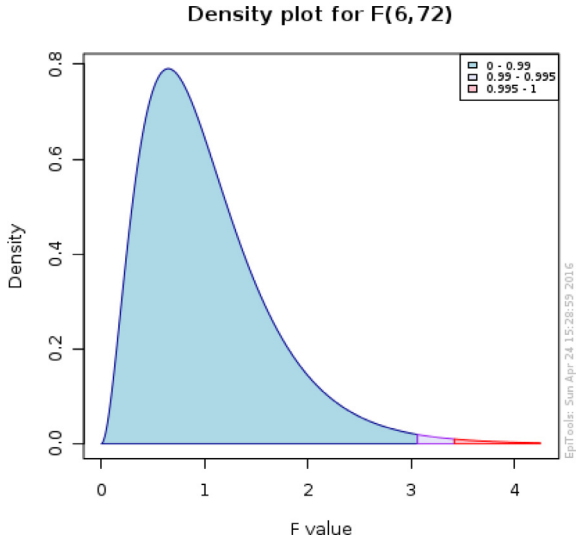

Fig. 14. Density Plot on the Degree of Freedom (6,72) based on the rate of error.

Table 12
Results of Friedman and Iman-Davenport test based on the rate of error.

Test	*Statistical value*	*Obtained critical value*	*Hypothesis*
Friedman	56.32	3.06	Rejected
Iman-Davenport	31.17	3.06	Rejected

Table 13
Results of Holm Test based on the rate of error (Control algorithm: CRO-FCM).

i	*Method*	*z-value*	*p-value*	$\alpha/(m-i)$	*Hypothesis*
6	***GA-K-means***	6.77	0	0.01	*Rejected*
5	***PSO-K-means***	4.84	6.492338E-7	0.005	*Rejected*
4	***GA-FCM***	4.66	0.000002	0.003	*Rejected*
3	***TLBO-K-means***	3.92	0.000044	0.0025	*Rejected*
2	***ETLBO-K-means***	2.73	0.003167	0.002	*Not Rejected*
1	***PSO-FCM***	2.01	0.022216	0.0016	*Not Rejected*

distance, number of iterations and error rate as compared to other methods in all datasets. Several statistical tests such as Friedman's rank test, Iman-Davenport test, Holm test were performed to verify the results and the proposed method is proved to be statistically significant.

In this paper, the proposed method is a hybrid approach based on chemical reactions for solving clustering problem. The literature shows many of the clustering models both in standalone form and hybrid form outperforms the classical clustering methods like K-means, K-medoid, FCM. But they suffer from some of the important issues like complex parameter tunings, large error rate etc. On the other hand, the method developed in this paper is able to tackle these issues by using a recently introduced chemical reaction based optimization algorithm. The simulation results of the proposed method have also shown that, it not only produces optimal results in less number of iterations, but also its average error rate is quite less as compared to other considered methods. Moreover, from the obtained fitness values, it can be clearly inferred that CRO-FCM outperforms the other methods in all the considered real world bench mark datasets. The advantages of proposed method are: (a) it is free from some complicated parameter tuning issues, (b) easily implementable due to simple structure, (iii) less error rate, (iv) statistically valid.

The method proposed in this paper for fuzzy clustering may be adapted to some other real life problems. The good simulation results obtained in this work encourage for extension of the method for these problems. Some of the important future directions include: (i) The adaptation of the proposed approach for handling more complex data, such as histograms, real forecasting data etc., (ii) The use of the proposed approach to train the higher order neural networks such as Pi-sigma network, functional link network etc., as standard PSO and FCM based methods have already been used to train the radial basis function networks (Tsekouras & Tsimikas, 2013). (iii) Also, the investigation of the proposed method may be done on solving some other data mining problems such as prediction, classification, forecasting. In addition to these, another future direction may comprise of solving some of the real life problems such as agricultural sectors, image processing, medical and health related problems etc.

References

Aghabozorgi, S., & Teh, Y. W. (2014). Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applications, 41*(4), 1301–1314.

Ahmadyfard, A., & Modares, H. (2008, August). Combining PSO and k-means to enhance data clustering. *Telecommunications, 2008. IST 2008.* international symposium on *(pp. 688-691)*. IEEE.

Alatas, B. (2011). ACROA: Artificial chemical reaction optimization algorithm for global optimization. *Expert Systems with Applications, 38*(10), 13170–13180.

Al-Mohair, H. K., Saleh, J. M., & Suandi, S. A. (2015). Hybrid human skin detection using neural network and K-means clustering technique. *Applied Soft Computing, 33*, 337–347.

Amjady, N. (2007). Short-term bus load forecasting of power systems by a new hybrid method. *IEEE Transactions on Power Systems, 22*(1), 333–341.

Asanambigai, V., & Sasikala, J. (2016). Adaptive chemical reaction based spatial fuzzy clustering for level set segmentation of medical images. *Ain Shams Engineering Journal.* doi:10.1016/j.asej.2016.08.003.

Bache, K., & Lichman, M. (2013). *UCI machine learning repository* [http://archive.ics.uci.edu/ml]. *School of information and computer science.* Irvine. CA: University of California.

Baral, A., & Behera, H. S. (2013). A novel chemical reaction-based clustering and its performance analysis. *International Journal of Business Intelligence and Data Mining, 8*(2), 184–198.

Bechikh, S., Chaabani, A., & Said, L. B. (2015). An efficient chemical reaction optimization algorithm for multiobjective optimization. *IEEE transactions on cybernetics, 45*(10), 2051–2064.

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms.* Springer Science & Business Media.

Carvalho, D. R., & Freitas, A. A. (2004). A hybrid decision tree/genetic algorithm method for data mining. *Information Sciences, 163*(1), 13–35.

Chen, B., Tai, P. C., Harrison, R., & Pan, Y. (2005, August). Novel hybrid hierarchical-K-means clustering method (HK-means) for microarray analysis. In *2005 IEEE computational systems bioinformatics conference-workshops (CSBW'05)* (pp. 105–108). IEEE.

Duan, H., & Gan, L. (2015). Elitist chemical reaction optimization for contour-based target recognition in aerial images. *IEEE Transactions on Geoscience and Remote Sensing, 53*(5), 2845–2859.

Duan, H., & Gan, L. (2015a). Orthogonal multiobjective chemical reaction optimization approach for the brushless DC motor design. *IEEE Transactions on Magnetics, 51*(1), 1–7.

Dunn†, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics, 4*(1), 95–104.

Dutta, S., Roy, P. K., & Nandi, D. (2016). Optimal location of STATCOM using chemical reaction optimization for reactive power dispatch problem. *Ain Shams Engineering Journal, 7*(1), 233–247.

Esnaf, Ş., & Küçükdeniz, T. (2009). A fuzzy clustering-based hybrid method for a multi-facility location problem. *Journal of Intelligent Manufacturing, 20*(2), 259–265.

Fister, I., Yang, X. S., & Brest, J. (2013). A comprehensive review of firefly algorithms. *Swarm and Evolutionary Computation, 13*, 34–46.

Hatamlou, A. (2013). Black hole: A new heuristic optimization approach for data clustering. *Information sciences, 222*, 175–184.

Izakian, H., & Abraham, A. (2011). Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. *Expert Systems with Applications, 38*(3), 1835–1838.

Jiang, H., Li, J., Yi, S., Wang, X., & Hu, X. (2011). A new hybrid method based on partitioning-based DBSCAN and ant clustering. *Expert Systems with Applications, 38*(8), 9373–9381.

Kanungo, D. P., Nayak, J., Naik, B., & Behera, H. S. (2016). Hybrid clustering using elitist teaching learning-based optimization: An improved hybrid approach of TLBO. *International Journal of Rough Sets and Data Analysis (IJRSDA), 3*(1), 1–19.

Lam, A. Y., & Li, V. O. (2010). Chemical-reaction-inspired metaheuristic for optimization. *IEEE Transactions on Evolutionary Computation, 14*(3), 381–399.

Lam, A. Y., Li, V. O., & Xu, J. (2013). On the convergence of chemical reaction optimization for combinatorial optimization. *IEEE Transactions on Evolutionary Computation, 17*(5), 605–620.

Li, J. Q., & Pan, Q. K. (2012). Chemical-reaction optimization for flexible job-shop scheduling problems with maintenance activity. *Applied Soft Computing, 12*(9), 2896–2912.

Li, Z., Nguyen, T. T., Chen, S., & Truong, T. K. (2015). A hybrid algorithm based on particle swarm and chemical reaction optimization for multi-object problems. *Applied Soft Computing, 35*, 525–540.

Lin, P. L., Huang, P. W., Kuo, C. H., & Lai, Y. H. (2014). A size-insensitive integrity-based fuzzy c-means method for data clustering. *Pattern Recognition, 47*(5), 2042–2056.

Naik, B., Nayak, J., & Behera, H. S. (2016). An efficient FLANN model with CRO-based gradient descent learning for classification. *International Journal of Business Information Systems, 21*(1), 73–116.

Nayak, J., Naik, B., & Behera, H. S. (2015). Fuzzy C-means (FCM) clustering algorithm: A decade review from 2000 to 2014. In *Computational intelligence in data mining-volume 2* (pp. 133–149). India: Springer.

Nayak, J., Naik, B., & Behera, H. S. (2015a). A novel chemical reaction optimization based higher order neural network (CRO-HONN) for nonlinear classification. *Ain Shams Engineering Journal, 6*(3), 1069–1091.

Nayak, J., Naik, B., Kanungo, D. P., & Behera, H. S. (2015b). An improved swarm based hybrid k-means clustering for optimal cluster centers. In *Information systems design and intelligent applications* (pp. 545–553). India: Springer.

Nayak, J., Kanungo, D. P., Naik, B., & Behera, H. S. (2016). Evolutionary improved swarm-based hybrid K-means algorithm for cluster analysis. In *Proceedings of the Second International Conference on Computer and Communication Technologies* (pp. 343–352). India: Springer.

Sarikprueck, P., Lee, W. J., Kulvanitchaiyanunt, A., Chen, V. C., & Rosenberger, J. (2015). Novel hybrid market price forecasting method with data clustering techniques for EV charging station application. *IEEE Transactions on Industry Applications, 51*(3), 1987–1996.

Satapathy, S. C., & Naik, A. (2014). Modified teaching–learning-based optimization algorithm for global numerical optimization—A comparative study. *Swarm and Evolutionary Computation, 16*, 28–37.

Shin, C. K., Yun, U. T., Kim, H. K., & Park, S. C. (2000). A hybrid approach of neural network and memory-based learning to data mining. *IEEE Transactions on Neural Networks, 11*(3), 637–646.

Silva Filho, T. M., Pimentel, B. A., Souza, R. M., & Oliveira, A. L. (2015). Hybrid methods for fuzzy clustering based on fuzzy c-means and improved particle swarm optimization. *Expert Systems with Applications, 42*(17), 6315–6328.

Song, W., Qiao, Y., Park, S. C., & Qian, X. (2015). A hybrid evolutionary computation approach with its application for optimizing text document clustering. *Expert Systems with Applications, 42*(5), 2517–2524.

Taşdemir, K., Milenov, P., & Tapsall, B. (2012). A hybrid method combining SOM-based clustering and object-based analysis for identifying land in good agricultural condition. *Computers and electronics in agriculture, 83*, 92–101.

Ting, T. O., Yang, X. S., Cheng, S., & Huang, K. (2015). Hybrid metaheuristic algorithms: Past, present, and future. In *Recent advances in swarm intelligence and evolutionary computation* (pp. 71–83). Springer International Publishing.

Truong, T. K., Li, K., & Xu, Y. (2013). Chemical reaction optimization with greedy strategy for the 0–1 knapsack problem. *Applied Soft Computing, 13*(4), 1774–1780.

Tsekouras, G. E., & Tsimikas, J. (2013). On training RBF neural networks using input–output fuzzy clustering and particle swarm optimization. *Fuzzy Sets and Systems, 221*, 65–89.

Wong, M. L., & Leung, K. S. (2004). An efficient data mining method for learning Bayesian networks using an evolutionary algorithm-based hybrid approach. *IEEE Transactions on Evolutionary Computation, 8*(4), 378–404.

Xia, M., Lu, W., Yang, J., Ma, Y., Yao, W., & Zheng, Z. (2015). A hybrid method based on extreme learning machine and k-nearest neighbor for cloud classification of ground-based visible cloud image. *Neurocomputing, 160*, 238–249.

Yang, F., Sun, T., & Zhang, C. (2009). An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization. *Expert Systems with Applications, 36*(6), 9847–9852.

Yang, Y., Huang, S., & Rao, N. (2008). An automatic hybrid method for retinal blood vessel extraction. *International Journal of Applied Mathematics and Computer Science, 18*(3), 399–407.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353. doi:10.1016/S0019-9958(65)90241-X.ISSN 0019-9958.