Journal of Computer Sciences

Original Research Paper

Fractional Lion Algorithm-An Optimization Algorithm for Data Clustering

[1]Satish Chander, [2]P. Vijaya and [3]Praveen Dhyani

[1,2]*Waljat College of Applied Sciences, P.O Box 197, P.C. 124, Rusayl, Muscat, Oman*
[3]*Banasthali University, Jaipur Campus, India*

Article history
Received: 13-04-2016
Revised: 03-08-2016
Accepted: 10-08-2016

Corresponding Author:
Satish Chander
Waljat College of Applied Sciences, P.O Box 197, P.C. 124, Rusayl, Muscat, Oman
Email: dimrisatish@gmail.com

Abstract: Clustering divides the data available as bulk into meaningful, useful groups (Clusters) without any prior knowledge about the data. Cluster analysis provides an abstraction from individual data objects to the clusters in which those objects reside. It is a key technique in the data mining and has become an important issue in many fields. This paper presents a novel Fractional Lion Algorithm (FLA) as an optimization methodology for the clustering problems. The proposed algorithm utilizes the lion's unique characteristics such as pride, laggardness exploitation, territorial defence and territorial take over. The Lion algorithm is modified with the fractional theory to search the cluster centroids. The proposed fractional lion algorithm estimates the centroids with the systematic initialization itself. Proposed methodology is a robust one, since the parameters utilized are insensitive and not problem dependent. The performance of the proposed rapid centroid estimation is evaluated using the cluster accuracy, jaccard coefficient and rand coefficient. The quality of this approach is evaluated on the benchmarked iris and wine data sets. On comparing with the particle swarm clustering algorithm, experimental results shows that the clustering accuracy of about 75% is achieved by the proposed algorithm.

Keywords: Fractional Lion Algorithm, Laggardness Rate, Sterility Rate, Rapid Centroid Estimation

Introduction

In Information technology, widespread use of the information leads to the amassing of the huge volume of information in many fields such as production, marketing, business etc. the bulk data must be grouped for the valid use. This leads to the development of the innovative methods to renovate the huge data into valuable information and knowledge. Data mining and Machine learning communities are the approach to meet the requirement of the data clustering (Yin *et al*., 2010). Clustering is useful for dividing large multidimensional data into distinguishable representative clusters (Yuwono *et al*., 2012). Clustering analysis has long played an important role in a wide variety of fields, whether for understanding or utility. In the context of clustering for understanding, clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes. The following are some examples: Biology, Information retrieval, climate, Psychology and medicine. In context of clustering for utility, cluster analysis is the study of techniques for finding the most representative cluster types. They are Summarization, Compression, Efficiently finding nearest neighbours etc (Kotsiantis and Pintelas, 2004; Lee and Olafsson, 2005). Cluster analysis groups data objects based only on the information that describes the object and their relationship. The goal is that the objects within a group be similar to one another and different from the objects in the other groups (Ji *et al*., 2012).

The Cluster analysis is performed using four steps. They are feature selection, Clustering algorithm, Cluster validation and Interpretation (Xu and Wunsch, 2005). The efficacy of the clustering application depends on the feature selection and extraction. The effective feature reduces the workload of the clustering procedure. The data sets subjected to clustering are initially selected based on the distinctive feature of the candidate set and the feature extraction uses few changes to cluster them into a meaning full candidate set from original data set. Second one is selection or design of clustering algorithm which is commonly integrated with the selection of corresponding proximity measure and construction of criterion function. Grouping of patterns are done by checking the similarity between them. The proximity measure explicitly has impact on the creation of the resultant clusters. Most of the clustering algorithms are directly or indirectly related to few characterization of proximity measure. Third one is cluster validation in which for a given a data set, a division is generated by the clustering algorithm, without considering the existence of the structure. Order of input patterns or the parameter identification may influence the results. Finally, results interpretation which is the ultimate goal of clustering. It provides significant insights from the original data to the users and so all issues are solved efficiently. Professionals in the respective domains infer the partitioning of data. Additional experimentations may be needed to assure the consistency of the knowledge obtained.

Several types of data clustering algorithm are available for the clustering. These algorithms are categorized into hierarchical (nested), partitional (unnested), exclusive, overlapping, fuzzy, complete and partial algorithms respectively. The requirements that a Clustering algorithm should satisfy are:

- Scalability
- Ability to deal with different types of attributes
- Easier input parameter determination with minimal domain knowledge
- Ability to deal with noise and outliers
- Interpretability and usability

The partitional clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. The K-means algorithm is the widely accepted technique as a partitional algorithm (Ji *et al*., 2012). If the clusters have the subclusters, then the hierarchical clustering type is achieved. In this, the set of clusters are organized as a tree or dendogram (Xu and Wunsch, 2005). Hierarchical algorithms are categorized as agglomerative methods and divisive methods (Xu and Wunsch, 2005). In over lapping clustering, the clustering reflects the fact that an object can simultaneously belong to more than one group where as in exclusive clustering; every objects are clustered as a single cluster. The fuzzy clustering clusters the object based on the member ship weight that is between 0 and 1. In other words, clusters are treated as fuzzy sets. Fuzzy C-means is the commonly used fuzzy clustering technique (Filho *et al*., 2015). As the name suggests, in the complete and partial clustering algorithm the object may or may not be included in the clusters.

There are many Evolutionary Algorithms (EAs) which draw inspiration from evolution by natural selection. Currently, there are several different types of EAs which include Genetic Algorithms (GAs), Genetic Programming (GP), Evolutionary Programming (EP) and Evolutionary Strategies (ES) (Lichman, 2013). The Fractional Lion algorithm proposed in this study is a population-based optimization algorithm that employs Survival of the Fittest (SF) model as the framework. One-dimension search is adopted in the paper and mutation to enhance the constringency speed and a new crossover strategy is adopted to avoid premature convergence. In addition to it, fractional lion theory is applied after the mutation to stabilize the centroid selection problem. The stabilization is done by the new solution point generated using the initialization itself. Thereby provide ease of search for the cluster centroids. The solution point out of the crossover, mutation and fractional lion enhances the rapid centroid estimation since the total solutions considered for the selection are higher in number.

The main contributions of the proposed paper are

New Optimization

A new optimization algorithm called FLA is proposed based on the lion algorithm (Rajakumar, 2012) for the optimization of the centroid estimation. The proposed algorithm is inspired by the pride lion behaviour. Since the behaviour is insensitive to the parameters, the proposed algorithm is more robust.

Clustering Process with Optimization

The clustering process with the optimized centroid point is done by making use of the fractional lion. The best possible solution points as the centroid are selected with the iteration till the tolerance is achieved. Thereby clustering process with the optimization is accomplished.

The rest of the paper is organized as follows. In section 2, some of the literature reviews regarding the centroid estimation using various algorithms are discussed. The motivation for the paper is conferred in section 3. Section 4 describes about the proposed algorithm. The result and discussion of the proposed paper is viewed in section 5. Finally, section 6 concludes this paper.

Literature Review

In this section, the literature review regarding the clustering algorithm for the centroid estimation is discussed accordingly for about 13 research papers.

MacQueen (1967), K-means algorithm (Ji *et al*., 2012) was developed and it was the first algorithm developed for clustering process. After the introduction of k-means algorithm, various algorithms were proposed in the literature for data clustering. Recently, (Huang *et al*., 2014) have proposed clustering algorithm by expanding the existing k-means-type algorithms by merging both intercluster separation and intracluster compactness. Initially, a set of novel objective functions for clustering was generated. Then, depending on the objective functions, the corresponding updating rules for the algorithms were logically developed. One of the popular techniques after k-means is fuzzy c-means clustering (Dunn, 1973). It considers the fuzzy theory for grouping the data objects. Recently, (Parker and Hall, 2014) have proposed two accelerated algorithms, such as Geometric Progressive Fuzzy C-Means (GOFCM) and Minimum Sample Estimate Random Fuzzy C-Means (MSERFCM) which apply statistical method to calculate the subsample size A common stopping criterion for accelerated clustering was established. The algorithms were evaluated with FCM and four accelerated variants of FCM. GOFCM's acceleration was four times higher than FCM and faster than SPFCM when applied on the six datasets used for experimentation.

Similarly, (Ji *et al*., 2012) have combined mean and fuzzy centroid to symbolize the prototype of a cluster and used co-occurrence of values based measure to examine the difference between data objects and prototypes of clusters. Further, the importance of various attributes towards the clustering process is also considered by this measure. Then, they presented an algorithm for mixed data clustering. At last, the proposed method was compared with conventional clustering algorithm by doing experiments on four real world datasets. PradiptaMaji (2011) has proposed an algorithm which depends on the theory of fuzzy-rough sets, which explicitly includes the details of sample categories in the clustering process. Here, the cluster development depends on sample categories. The efficacy of the FRSAC algorithm and the comparison with the available supervised and unsupervised gene selection and clustering algorithms is verified on six data sets based on the class separability index and predictive accuracy of the naive Bayes' class.

Yin *et al*. (2010) has proposed an adaptive Semi-supervised Clustering Kernel Method depending on Metric learning (SCKMM) to solve the issues stated above. Particularly, they initially an objective function is developed from pair wise constraints for automatic extraction of Gaussian kernel parameter. Then, they used pair wise constraint-based K-means technique to deal the constraints violation problem in data clustering. Also, they introduced metric learning into nonlinear semi-supervised clustering to enhance separability of the data for clustering.

Some of researchers formulated the clustering problem as optimal search problem and they have applied optimization algorithm for data clustering. Accordingly, Yuwono *et al*. (2014) have proposed an algorithm called, PSC for solving clustering problems. PSC algorithm for substitution was constructed such that it is easy and efficient to implement. An approach, called particle reset and white noise update was introduced. This implementation was called as Rapid Centroid Estimation (RCE). The update rules of PSC are simplified by RCE and it reduced the computational complexity by improving the effectiveness of the particle trajectories. Also, Binu (2015) has proposed three newly designed objective functions along with four existing objective functions with the help of optimization algorithms like, genetic algorithm, cuckoo search and particle swarm optimization algorithm. Here, three different objective functions were designed including the cumulative summation of fuzzy membership and distance value with normal data space, kernel space as well as multiple kernel space. In addition to the existing seven objective functions, totally, 21 different clustering algorithms were discussed and the performance was validated with 16 different datasets which are synthetic, small and large scale real data.

Sheng *et al*. (2014) have used three local searches of various features to effectively utilize the decision space. Moreover, they developed an adaptive niching method, in which its parameter value is regulated dynamically based on the occurrence of the problem and the search progress and it is incorporated into the algorithm. The adaptation strategy was depends on a formulated population diversity index and is used to improve fitness and genetic diversity. As a result, diverged niches of high fitness can be created and preserved in the population which makes the technique more suitable for efficient investigation of the complex decision space of clustering problems. The consequent algorithm is used to optimize a consensus clustering criterion, which is recommended to achieve reliable solutions. To assess the algorithm, series of experiments were carried out on real and synthetic data and evaluated with existing methods.

Bandyopadhyay (2011) have presented a combination of a measure for cluster validity and the concept of stability to describe two objective functions which are concurrently optimized for clustering. Although the mean value of the cluster validity index, calculated over various bootstrapped samples of the data was considered as the primary goal, its average difference was considered as the next goal to be optimized. The later indicates the stability of the partitioning of the different bootstrapped samples of the data. An algorithm called AMOSA was adopted as the fundamental optimization method. A semi supervised approach was recommended for finding a capable solution from the set of Pareto-optimal solutions.

Kiranyaz *et al*. (2010) have proposed two methods that effectively concentrate on various severe issues in the area of Particle Swarm Optimization (PSO). Multidimensional (MD) PSO improves the native structure of swarm particles so that inter-dimensional passes with a dedicated dimensional PSO process can be made. Hence, in an MD search space, the optimum dimension is unidentified and swarm particles try to find the positional and dimensional optima. In order to enhance the searching performance of clustering, authors tried to combine two different search algorithms. Accordingly, Kuo *et al*. (2012) have proposed a dynamic clustering technique depending on Particle Swarm Optimization (PSO) and genetic algorithm (GA) (DCPG) algorithm. The DCPG algorithm clusters data automatically by investigating the data without specific number of clusters. The computational results of four benchmark data sets indicates that the DCPG algorithm have superior stability and validity compared with the dynamic clustering approach based on binary-PSO (DCPSO) and the Dynamic Clustering approach based on GA (DCGA) algorithms. Moreover, the DCPG algorithm was employed in clustering the Bills of Material (BOM) for the Advantech Company in Taiwan.

Motivation

The motivation of the proposed algorithm is concerned with the adaptive centroid estimation for the bulk data sets. In this section, the problem definition and the challenges responsible for the evolution of the proposed FLA are discussed.

Problem Definition

Let as assume that the input data base is X. The data base contains y number of the objects. Each of the data object is represented with the z attributes. For example data base X is represented as $X = \{X_1, X_2,\ldots,X_y\}$; $1 \leq i \leq y$. Every data object within the data base is represented along with the attributes. The attribute indication of the data base is, $X_j = \{x_{j1}, x_{j2},\ldots,x_{jz}\}$; $1 \leq j \leq z$. The ultimate challenge is to cluster the data objects in the data base into k clusters using the clustering algorithms. The clustering over the input database can be signified as the identification of k centroids which are represented as, $W = \{W_1, W_2,\ldots, W_k\}$; $1 \leq l \leq k$. Here, every centroid is represented with z attribute values as like, $W_l = \{W_{l1}, W_{l2},\ldots, W_{lz}\}$. The centroid selected must be checked out for the fitness evaluation for choosing the optimal centroid to group the cluster.

Challenges

Clustering finds a challenge of searching the optimal centroids which should be optimum to divide the data into k number of partition. So, clustering problem can be formulated as optimal searching problem. It can be stated that k number of centroids should be found out from the data space provided for the input data. Recently, the clustering searching problem is solved in (Yuwono *et al*., 2014) using Particle Swarm Clustering (PSC). In PSC, the centroid estimation was done using the position updating formula developed by them. Again, the evaluation of every centroid is done using the sum of squared distance. When analysing the PSC algorithm, these are challenges are identified to further extend the work:

- PSC has the risk to converge in local optimal solutions (or) clusters due to arbitrary assignment of weights
- The particle position updation does not comprise the data characteristics to initiate the cluster centroids so this may become complex because of wide data distribution, time series characteristics and high dimension
- It aims to find the global centroids throughout the process, rather than focusing on initialization part
- The termination strategy has not made the converging procedure to be aware of the quality improvement of centroids
- As per (Binu, 2015), the algorithmic effectiveness is decided by objective function but this work (Yuwono *et al*., 2014) utilizes the Sum of Squared Distance (SSD) as objective function even though a lot of improved objective functions are presented in the literature
- Also, data space-based objective function affect the convergence performance based on the characteristics of datasets such as, range of values, dimension, image and data type (integer or floating point)

Proposed Algorithm: Fractional Lion Algorithm for Rapid Centroid Estimation

In this section, the proposed fractional lion algorithm is discussed decoratively. Figure 1 shows the overall block diagram of the Proposed Fractional Lion Algorithm (FLA).

Initially, the data sets subjected to grouping are accepted as the input for the clustering algorithm. The data selection is based on the object vs. attribute. The object based on the attribute is x_{j1} (consideration). Here j is the attribute upon which the object is selected. The selected object points are then subjected, to the clustering algorithm. The clustering algorithm used here is the fractional lion algorithm. Before starting the solution point estimation for the process, initiation of the solution point is necessary.

The objects are randomly selected as the solution point within the interval for the clustering. It may be the first two objects or elected within the data set. In the fractional lion, primarily the solution points are initialized randomly and then they are imperilled to be the solution constraints. Based on the selected vectors of the solution point, cross over and mutation is done. Single point cross over with dual probability is used for the crossover of the solution vector points. The crossover is done by interchanging the points in the vectors within the clustering range. The clustering gives away new solution points. On the crossover solution points, mutation is done. The mutation gives away the new solution points. Rapid mutation is done to obtain the solution vectors. Then the fitness evaluation on the obtained solution vectors chooses the solution points possibility to present in the clustering process. The steps followed hereby are similar as that of the Lion algorithm. The modification took upon after the mutation.

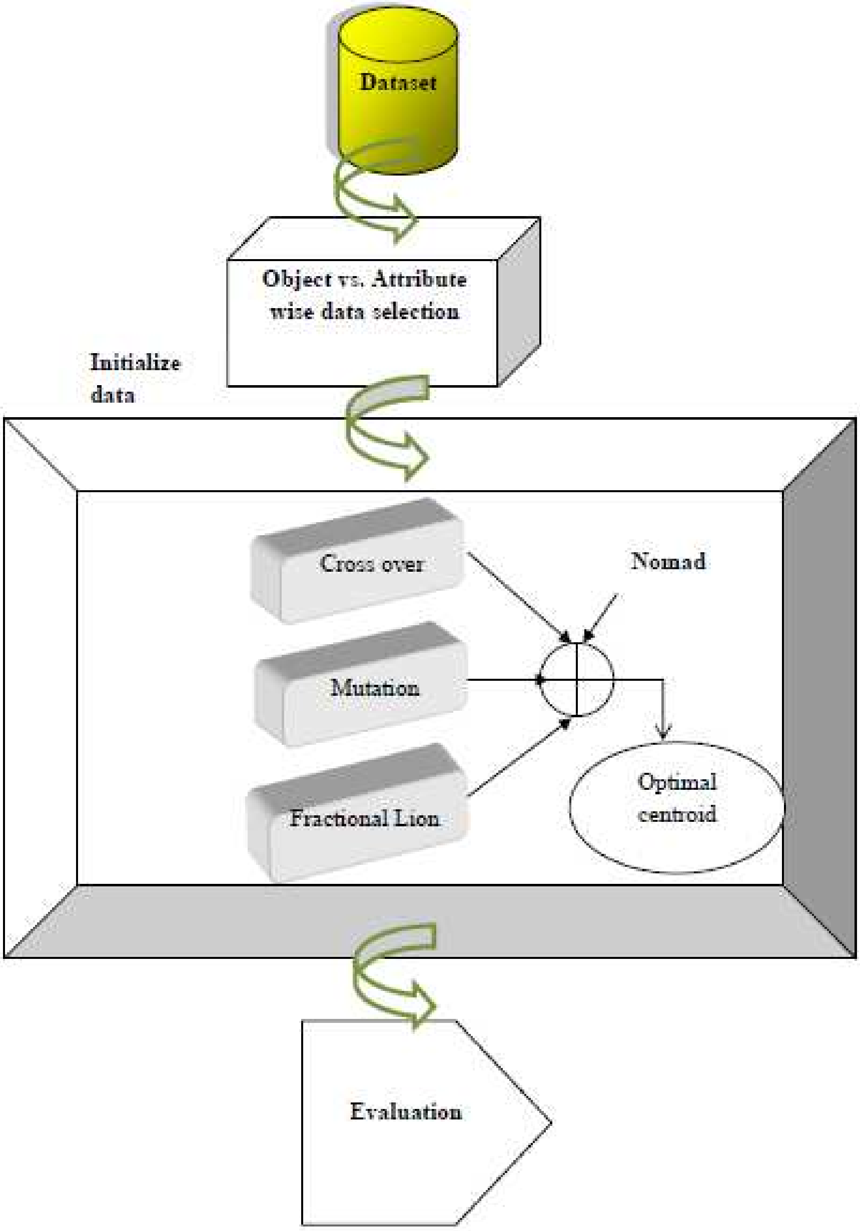


Fig. 1. Block diagram of proposed FLA algorithm

In the FLA, the lion algorithm is modified with the fractional theory. The new solution vectors as the solution points are generated in addition to the solution point out of the mutation and cross over. The wandering solution point within the range, if exceeds the fitness value of the initialized solution vector, the roving solution point takes the random points place. The solution point mentioned here are the centroid points. Then using the fitness function, the best optimal centroid for the clustering is selected from the calculated solution points and iterated until the tolerance is achieved.

After the centroid point estimation i.e., optimal centroid point, grouping is done based on it. The proposed algorithm because of the availability of the bulk solution point helps in the rapid centroid estimation. Hereby the proposed algorithm will be useful for data clustering with the bulk data sets, different scale etc. In the overall block diagram provided in the Fig. 1, nomad indicates the roving solution point which takes place of the initiated solution point if the fitness value exceeds it. The evaluation block in the diagram, corresponds the evaluation of the centroid point selected with the performance metrics which may be of internal metrics or the external metrics.

```
**FLA Algorithm**
**Input:** Initial Point $X^M, X^F$ & $X^N$
**Output:** Best matching point $X^{best}$
**Procedure**
**Start**
Read initial point, $X^M, X^F$
Calculate the fitness function
    Fertility evaluation
        If $f^{ref} \leq f(X^M)$
$f^{ref} \leftarrow f(X^M)$
        End
If $f(X^{F+}) < f(X^F)$
$X^F \leftarrow X^{F+}$
        End
Reset $L_r$ *and* $S_r$
Cross over and Mutation $X^C$ and $X^{New}$
Gender clustering $X^{M\_C}$ *and* $X^{F\_C}$
Set Ac
**Fractional based generation (** Assume $X_l^M$ *and* $X_{l+1}^M$ are same)
$X^{LION} = \alpha X_l^m + \frac{1}{2}\alpha X_{l-1}^M$
    Territorial defence
        If $X^N$ Wins
$X^M \leftarrow X^N$
        End
    Territorial take over
*if* $f(X^M) > f(X^{M\_C})$
$X^M = X^{M\_C}$
*else if* $f(X^F) > f(X^{F\_C})$
  $X^F = X^{F\_C}$
End
*clear* $S_r$
    Iterate until;
$N_f > N_f^{\max}$ ; $X^{best}$ attained.
**END**
```

Fig. 2. Pseudo code of the proposed Fractional Lion algorithm

Fractional Lion Algorithm

The fractional algorithm proposed for the rapid estimation of the centroid is discussed in this section. This paper makes an attempt to introduce a novel optimization algorithm, with modified fractional theory called the Fractional Lion (FLA) algorithm. The proposed algorithm is based on Lion Algorithm (Rajakumar, 2012), which is based on lion's unique social behaviour. Survival of the Fittest (SF) is the main idea. For a lion pride, only the few strongest male lions can remain in the pride. Thus, SF model is adopted in the algorithm. In the SF model, three main evolution strategies are developed according to the lion pride behaviour: (1) Two best members occupy all the mating resources of the pride; (2) the best member of the new offspring is trained to be stronger; (3) evolution stagnation leads to the takeover by new individuals and long evolution stagnation leads to a mutation to the best member. The fractional lion algorithm is utilized for the rapid estimation of the cluster centroids. Optimization, which is a process of seeking optima in a search space, is analogous to the evolution process (the best member represents the optima obtained) of animals in nature are achieved by the modified FLA.

The pseudo code for the proposed fractional Lion algorithm is shown in Fig. 2.

Step 1: Pride generation and subject to solution constraints.

Let X^M, X^F and X^N be the pride generation of the lions considered for the FLA. Here M denotes the male lion, F denotes the female lion and N denotes the nomad lion. The value of the X^M and X^F are given by:

$$X^M = \left[x_1^m, x_2^m, \ldots\ldots\ldots\ldots, x_L^m\right]$$

$$X^F = \left[x_1^F, x_2^F, \ldots\ldots\ldots\ldots\ldots, x_L^F\right]$$

where, L is length of the solution vector.

After the initiation, the fitness of X^M, X^F and X^N are calculated using the fitness equation i.e., Mean Square error.

The fitness value of the all the lions are stored. The reference fitness is set as the fitness value of the male lion.

Step 2: Fertility evaluation

This stage evaluates and ensures the fertility of the territorial lion and lioness. By doing so, the converging in the local optima is avoided.

Here the factors considered for the evaluation are listed below:

f^{ref} the Reference fitness function of X^M, L_r is Laggardness rate, S_r is Sterility rate, w_c is the Female update count, q_c is the Female generation count, X^{F+} is the Updated female lion.

The value of the laggardness rate and the sterility rate are chosen irrespective of gender of the lions. The values of S_r and L_r are set from the biological motivation.

When the algorithm begins, the value of S_r and L_r are initialized as zero. During the fertility evaluation, the values get incremented. In pseudo code, primarily the male lions are evaluated based on the laggardness rate and female lions are evaluated based on the sterility rate. The generation count of the female lion is set to be around 10 based on the trial and error method.

Calculation of X^{F+}:

$$X_l^{F+} = \begin{Bmatrix} x_k^{F+}; if\ l = k \\ x_k^F; otherwise \end{Bmatrix}$$

$$X_k^{F+} = \min\left[x_k^{\max}, \max(x_k^{\min}, \nabla_k\right]$$

$$\nabla_k = \left[x_k^F + (0.1r_{r2} - 0.05)\,(\,x_k^M - r_1 x_k^F\right]$$

Here l and k are the vector elements of the solution vector L. They are random integers generated within the interval $[0, L]$. ∇_k is the female update count function, r_1 and r_2 are the random integers generated within the interval [0,1].

Step 4: Cross over and mutation

The cross over and mutation are the significant operators for the evolutionary optimization. The maximum littering rates of four cubs are followed in the cross over.

Cross Over

The cross over used here is the single point cross over with the dual probabilities with the random cross over probability as C_r.

Cross over operation is given as:

$$X^C(R) = B_R \circ X^M + \overline{B_R} \circ X^F; R = 1,2,3,4$$

Where:
X^C = Represents the Cub obtained from the cross over
R = Cross over mask length, the values are 1' and 0's based on the C_r
° = Hadamord product

Mutation

X^C are then subjected to mutation to form X^{New}. The mutation is done with the mutation probability T_r. The obtained X^C and X^{New} are placed in a cub pool and then subjected to gender clustering.

Gender Clustering

The gender clustering is done to extract the male and the female cubs separately. Based on the physical nature and the cubs with the first and second fitness are selected as X^{M_C} and X^{F_C} respectively.

M_C denotes the male cub and F_C denotes the female cub.

Step 5: Cub Growth Function

The cub growth function is the mutation function. The extracted male and female cubs from the gender clustering are subjected to the mutation. If the mutated cubs have greater fitness value, then the mutated cubs are considered as the new cub male and female.

A_c is the age of cub which is incremented after mutation to illustrate the growth of the cub towards the maturity. N_r is the growth rate which must not exceed the mutation rate T_r.

Step 6: Fractional calculus-based generation

To increase the centroid estimation speed, i.e., for rapid estimation, a new male lion is generated using the fractional line theory.

For l^{th} element of the solution vector L.

Calculate the fitness function; if the fitness function is same or even if not same, the fractional theory is applied to get the new male lion. The lion algorithm is modified with mathematical theory called, Fractional Calculus (FC) (Pires *et al.*, 2010). The function α of the fractional theory is a constant value.

The fractional calculus based generation is done using the equation given below:
For the fitness value:

$$f\left(X^M_{l+l}\right) = f\left(X^M_l\right)$$

If no change in the best solution:

$$X^M_{l+1} = X^M_l$$

Rearranged in order to modify the order the solution derivative:

$$X^M_{l+1} - X^M_l = 0$$

Discrete version of the derivative of order $\alpha = 1$ for X^M_{l+1} leads to following expressions as:

$$D^\alpha\left[X^M_{l+1}\right] = 0$$

For 0≤α≤1; with order 2, discrete version can be elaborated as:

$$D^\alpha\left[X^M_{l+1}\right] = X^M_{l+1} - \alpha X^M_l - \frac{1}{2}\alpha X^M_{l-1}$$
$$X^M_{l+1} - \alpha X^M_l - \frac{1}{2}\alpha X^M_{l-1} = 0$$
$$X^M_{l+1} = \alpha X^M_l + \frac{1}{2}\alpha X^M_{l-1}$$
$$X^{Lion} = X^M_{l+1}$$
$$X^{Lion} = \alpha X^M_l + \frac{1}{2}\alpha X^M_{l-1}$$

where function α is a constant value.

Step 7: Territorial defence

The territorial is the primary operator of the lion to increase the search space in a wider way. The nomad lion does the survival fight with the territorial lion. If they win, then the pride and nomad coalition updates the territorial lion. The nomad lions taken here are X^N_l and X^N_2. The initialization is based on the laggardness rate since the only the male lions are considered for the territorial defence. The two lions are considered and the survival fight is done between the nomad lion with the greater fitness function. If it fails to survive, then the next nomad lion is updated and the survival fight goes on. The updation of the nomad lion happens only if the following condition is satisfied.

If X_2^N is to be updated because of the loss in Fight of X_1^N, then $E_2^N \geq e$. The value of E_2^N is calculated using:

$$E_2^N = \exp\left(\frac{d_2}{\max(d_1, d_2)}\right)\frac{\max\left(f\left(X_1^N\right), f\left(X_2^N\right)\right)}{f\left(X_2^N\right)}$$

where, d_1 is the Euclidean distance between X_1^N and X^M. d_2 Is the Euclidean distance between X_2^N and X^M.

Step 8: Territorial takeover

The territorial take over takes place when the cub reaches the maximum age. It is concerned with the curb growth function. It is the process of giving the territory to cub lion and lioness after they mature and become stronger than the male and female lion. Once the territorial takeover is done, the value of the sterility rate is set zero. After that, one generation is completed. The value of the generation count is incremented to one.

Step 9: Iteration

The proposed fractional lion algorithm is iterated until:

$$N_f > N_f^{\max} \text{ is achieved}$$

where, N_f is the number of function evaluations.

Adapting Clustering Process to Solve with Search Algorithm

In this section, the clustering process done along with the search algorithm is discussed. The search algorithm used in the proposed algorithm is the fractional Lion algorithm. The search algorithm searches the points which are eligible to be as the centroid point. Among the centroid point estimated, the best centroid point is chosen upon the functional evaluation and the if the best centroid is attained, it will replaces the original centroid point randomly selected at the time of initiation and then the searching algorithm is iterated. Then again, the centroid points are selected based on the function value and iterated till the tolerance is achieved. The solution encoding and the fitness function of the proposed FLA are discussed in this section.

Solution Encoding

In this study, we adopted an encoding scheme based on the proposed fractional lion algorithm. The input data base is X with y object. If $y = 10$, then:

$$X = \{x_1, x_2 .... x_{10}\}$$

If the attribute z is 2, then for grouping the data into the cluster, the value of K must be adapted. If the value of K is 5, then the total number of element in a single group is $[K\times z]\rightarrow[5\times2] = 0$.

The value of the $K_{\max}$ is defined by the user.

The centroid point is selected randomly from the data point within the interval for which the efficient grouping is done (Fig. 3).

Fitness Evaluation

The fitness evaluation chooses the best centroid point for the clustering of the objects.

Let W_1 and W_2 be the two centroid points taken into consideration for the fitness evaluation in the data sets in the data sets X. The solution here is to choose the best centroid for the clustering. It is estimated by calculating the mean distance between the centroid point and the objects which are to be clustered.

Figure 4, shows the fitness evaluation procedure. The distance for the individual object from the centroid points must be calculated first. Then the mean difference between the distances is calculated to obtain the best possible centroid point to grouping.

I.e., For d_1, d_2, d_3, d_4 shown in the Fig. 4. Fitness evaluation for the centroid point is given by:

$$F = \sum_{i=1}^{k}\sum_{\substack{j=1 \\ j\in i}}^{n_i} \| D_j - W_i \|$$

where, D is the distance for the data point in the relevant cluster and W is the centroid point for K clusters.

The net distance D due to the object of the individual cluster is given by:

$$D = \left(\sum_{l=1}^{4} | x_{lz} - x_{ly} |^2\right)$$

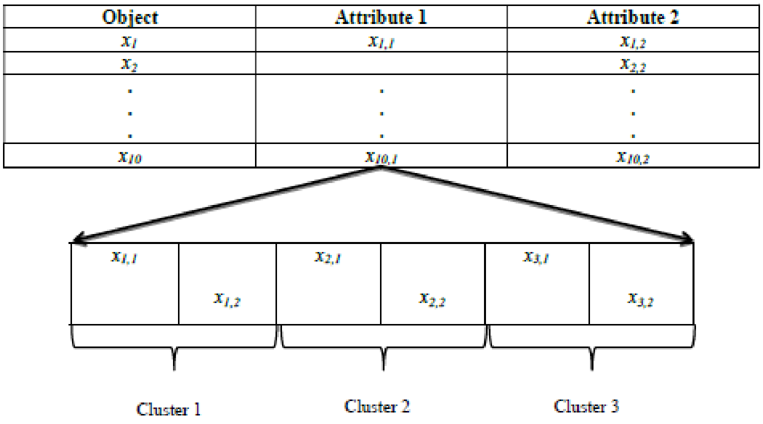
The Euclidean distance formulas for the object $x1$ with attribute z and y:

$$d = \sqrt{(x_{1z} - x_{1y})^2 + (x_{2z} - x_{2y})^2}$$

From the fitness function, the best centroid points for the individual data point are selected.

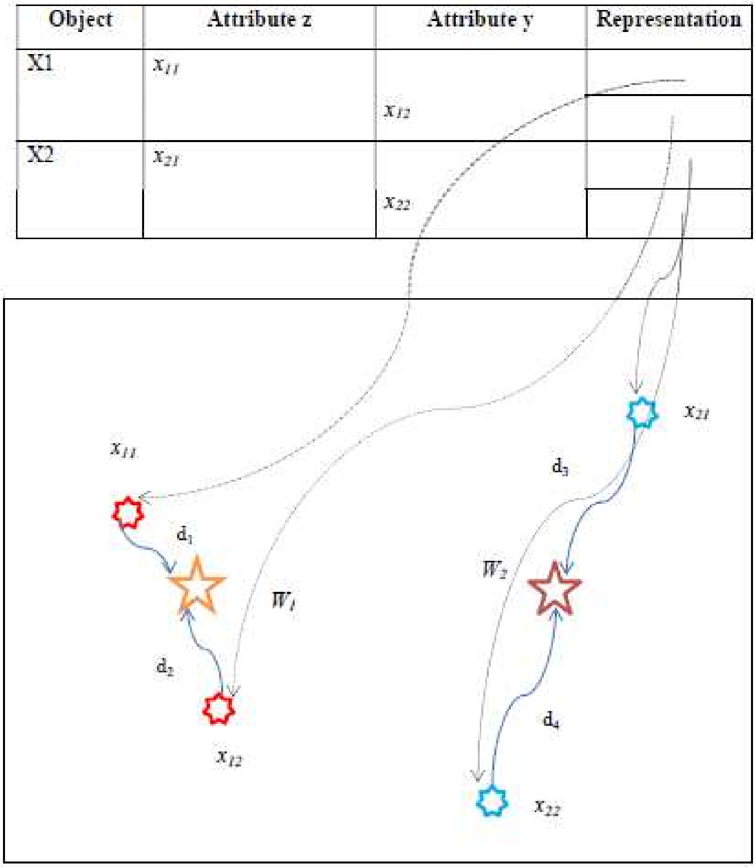
Clustfractlion: A New Clustering Algorithm

The Clustfractlion is the proposed algorithm for the rapid estimation of the centroid point. The proposed FLA is based on the lion algorithm. The lion algorithm utilizes the unique behaviour of the lion for the solution point estimation. For parental lions, four cubs from the cross over and four cubs from the mutation, one from the fractional lion calculus generation and a nomad lion are obtained. From the population, the best lion is selected based on the fitness function. The same procedure is used for the data clustering.



Object	Attribute 1	Attribute 2
x_1	$x_{1,1}$	$x_{1,2}$
x_2		$x_{2,2}$
. . .	. . .	. . .
x_{10}	$x_{10,1}$	$x_{10,2}$

Fig. 3. Solution encoding of the proposed fractional lion algorithm



Object	Attribute z	Attribute y	Representation
X1	x_{11}	x_{12}	
X2	x_{21}	x_{22}	

Fig. 4. Fitness evaluation

If X is the data set which is subjected to the clustering, the centroid points are selected from the interval with in which the grouping is done. The initial centroid points are taken as consideration and given into the FLA. The fractional algorithm, estimates the solution point i.e., centroid point. The solution points initialized are subjected to the cross over and mutation as like that of the lion algorithm. Once the new points are generated, the fractional calculus based solution point is generated based on the condition of the random solution points. Thus many solution points are generated out of the algorithm. Upon the calculated solution point, the fitness evaluation is done using the mean square error. After the fitness calculation, the centroid point with the optimal fitness is considered as the centroid point and the process of FLA is iterated until the tolerance is achieved.

The groupings of the objects are done using the estimated optimal centroid point. Proposed algorithm makes the clustering easier for the data sets with bulk data point by increasing the solution point.

Results and Discussion

In this section, the results and discussion of the proposed FLA is discoursed. The performance of the proposed algorithm is evaluated based on the internal and external evaluation metrics.

Experimental Set Up

Tools and Existing Works

The experimentation of the FLA is done using Personal Computer with following specification:

- Intel core i3 processor
- 2 GB RAM
- Windows 10 OS

And the software tool needed for the experimentation is:

- JAVA version 8

The characteristic of the proposed algorithm was exposed by comparing with the existing algorithm like Particle Swarm Clustering (PSC) (Yuwono *et al*., 2014), modified Particle Swarm Clustering (mPSC) (Wan *et al*., 2012) and Lion algorithm (Rajakumar, 2012). The result of the proposed algorithm over the bench marked data sets are viewed in experimental results.

Datasets

The performance of the proposed algorithm was benchmarked using two openly available data sets. The data sets used here are iris and wine (Lichman, 2013).

Datasets Description

Iris Dataset

Iris data set is the best known data base for the pattern recognition. It consist of three classes with 50 instances each. The attributes concerned with the iris data set is 4 which is of numeric type. Its characteristic is multivariate with real attributes.

Wine Dataset

Wine dataset is distinguished from the iris because of the well behaved class structure. The total numbers of instances are 178 i.e., in class I 59 instances, in class II 71 instances and in class III 48 instances. The attributes concerned with the wine datasets are continuous and 13 in total.

Evaluation Metrics

The evaluation metrics taken into consideration for the experimentation is enumerated in this section. The evaluation metrics which are considered in the experimentation are of two types. They are internal metrics and external metrics.

Internal Evaluation Metrics

Mean square error is the internal metrics by which the evaluation over the datasets is done to prove the efficiency of the proposed FLA algorithm over the existing data clustering algorithm.

Mean Square Error

The MSE value must be minima for the better clustering. The MSE is the fitness calculation formula which is newly developed in the proposed algorithm. The mean square error is calculated between the centroid points in the clustering. The point with the minimum error is accepted as the optimal centroid. The mean square error is given by:

$$MSE = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \in i}}^{n_i} \| D_j - W_i \|$$

where, D is the sum of the Euclidean distance in the data points of the individual cluster and W is the centroid point.

External Evaluation Metrics

The external evaluation metrics taken into consideration for the performance evaluation are rand coefficient, jaccard coefficient and clustering accuracy. The jaccard and rand are the similarity coefficient.

Clustering Accuracy

The clustering accuracy is given by:

$$CA = \frac{\left(\sum_{i=1}^{K} \max_{j-\{1,2,\ldots,K\}} \left\{ 2 \frac{\left| C_i \cap P_{mj} \right|}{\left| C_i + \left| P_{mj} \right| \right|} \right\} \right)}{K}$$

Here, $C = \{C_1,\ldots,C_K\}$ is a labelled data set that offers the ground truth and $P_m = \{P_{m1},\ldots, P_{mK}\}$ is a partition produced by a clustering algorithm for the data set.

Rand Coefficient

The rand coefficient is given by:

$$\text{Rand co} - \text{efficient}, RC = (SS + DD) / (SS + SD + DS + DD)$$

Here, *SS*, *SD*, *DS*, *DD* represent the number of possible pairs of data points where:

- *SS*: Both the data points belong to the same cluster and same group
- *SD*: Both the data points belong to the same cluster but different groups
- *DS*: Both the data points belong to different clusters but same group
- *DD*: Both the data points belong to different clusters and different groups

Jaccard Coefficient

The jaccard coefficient is given by:

$$\text{Jaccard co}-\text{efficient, } JC = (CC)/(CC + CE + EC)$$

Here, *CC*, *CE*, *EC* represent the number of possible pairs of data points where:

- *CC*: Both the data points belong to the same cluster and same group
- *CE*: Both the data points belong to the same cluster but different groups
- *EC*: Both the data points belong to different clusters but same group

Experimental Results

The experimental results of the proposed FLA algorithm over the existing algorithms are viewed in this section. The performance curve is plotted for the evaluation metrics. The experimentation is done using two cluster sizes to enumerate the efficiency.

Convergence Analysis

The convergence analysis of the fractional line algorithm over the data sets iris and wine are discussed in this section. For the cluster 2 and 3, the convergence analysis is analysed with the help of the total number of iteration and the corresponding fitness function. The analysis curve is plotted between the number of iteration and fitness value.

For the cluster 2, Fig. 5 shows the convergence analysis of the iris data set. On increase in the iteration, the value of the fitness function reduces for determining the optimum local minimum value. For the first iteration in the cluster 2, the fitness value out of the PSC algorithm is 398.16. The modified PSC which is a modification of the particle swarm optimizer have increased fitness value of 342.62 at the iteration 1 whereas the lion algorithm has further reduced value of 308.39. The proposed algorithms with the fractional theory have the value obtained as 304.18. The value of the fitness function is low for the Fractional lion algorithm compared to the previous search algorithm. The fitness functions have a drastic change when considering the PSC and the fractional Lion. In the following iterations, the value of the fitness function due to fractional lion is considerably lower compared to the existing algorithms.

The fitness value for the wine data set is shown in the Fig. 6. For the second cluster, the difference in the iteration and fitness value are discussed. The characteristic of the fitness function is lower in the proposed algorithm compared to the existing algorithm. For the first iteration, the fitness value is 1100 for PSC algorithm. The value is same for all the other exiting algorithm and proposed system i.e., 1000. But the effect happens, with increase in the iteration. From 1100, the fitness value reaches 1000 but proposed methodology reduces to about 726.5. The reduction in the fitness value is the most significant factor in the proposed algorithm.

At iteration 10, the value of the mPSC and Lion algorithm are 987 and 737 respectively. Even though, they have effect on the PSC algorithm, the efficacy is not effective over the improved fractional Lion algorithm. For the cluster size of 2, the proposed algorithm is most affective for the clustering with the minimal fitness value of 726.

The following two performance curve represents the convergence analysis of the proposed fractional lion algorithm for the cluster of size 3. Figure 7 and 8 shows the convergence analysis curve for the open data sets iris and wine respectively with cluster size 3.

In Fig. 7, for the iris data set the optimal fitness value obtained is 133 at the 10th iteration. From the analysis curve, the value of the fitness for the PSC, mPSC and lion algorithm are 204, 196 and 156.92 respectively. On comparison, the increased iteration results in the decrement of the fitness value in all the algorithms. But the rate at which the value gets decreased is completely dependent on the increased search point. The proposed algorithms have the increased search point over the existing algorithm and hence it is more effective for the clustering.

Figure 8 shows the convergence analysis curve of the wine data set. Here the optimal value of the fitness achieved by the proposed algorithm at the 10th iteration is 133. The fitness value decrease about 100 compared to the 10th iteration of the cluster size 2. At the 3rd iteration, the fitness value for the PSC algorithm is 305.56 where as for mPSC and lion algorithm, 271.49 and 201.38 respectively. The value decreased for about 70% compared to the PSC algorithm in the proposed algorithm with the value of 198.37.

On the convergence analysis of the proposed Fractional lion algorithm over the existing algorithms like PSC, mPSC and Lion algorithm, the fact which is evident from the efficient is evident from the analysed fitness value is that the proposed algorithm is more effective for the clustering with the rapid centroid estimation.

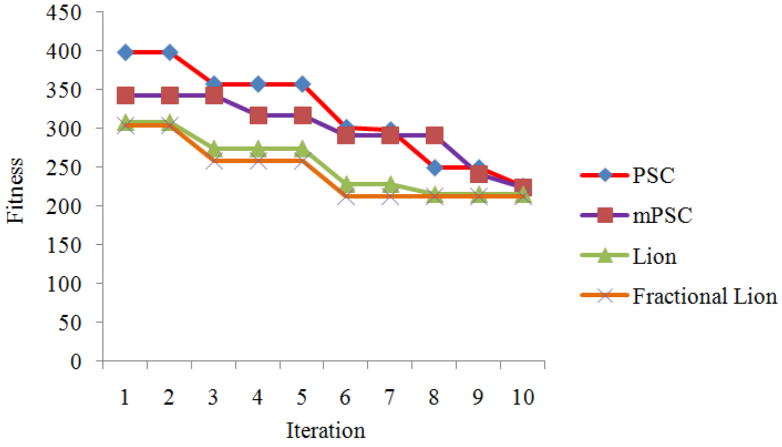


Fig. 5. Convergence analysis of Iris datasets cluster 2

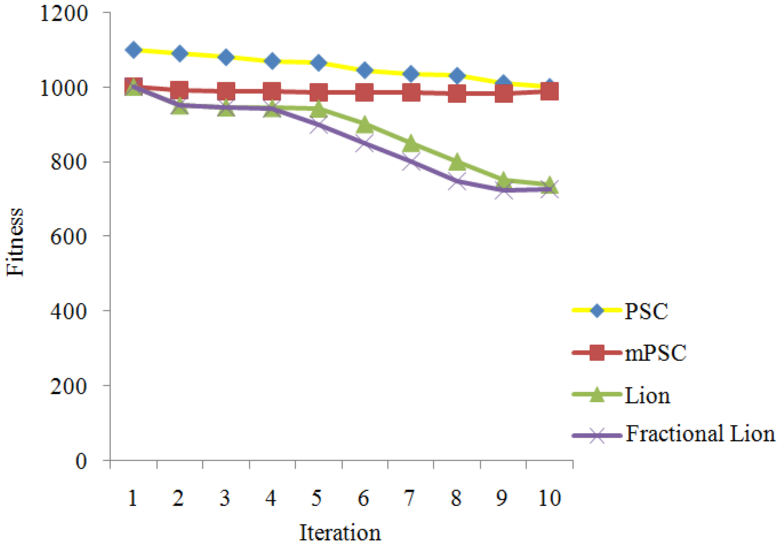


Fig. 6. Convergence analysis of Wine datasets cluster 2

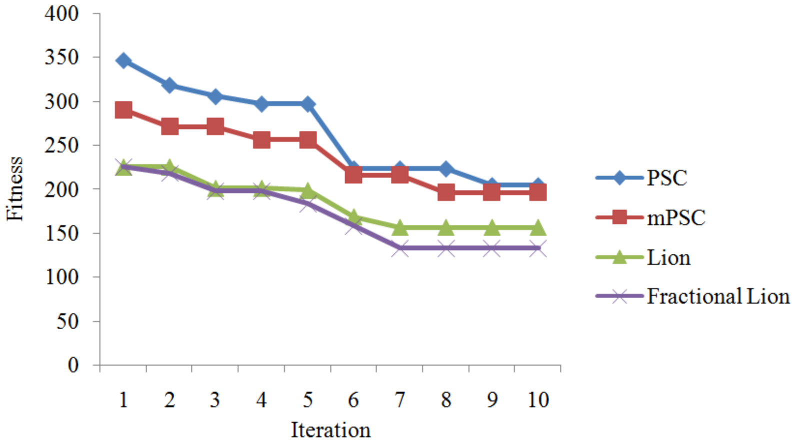


Fig. 7. Convergence analysis of Iris datasets cluster 3

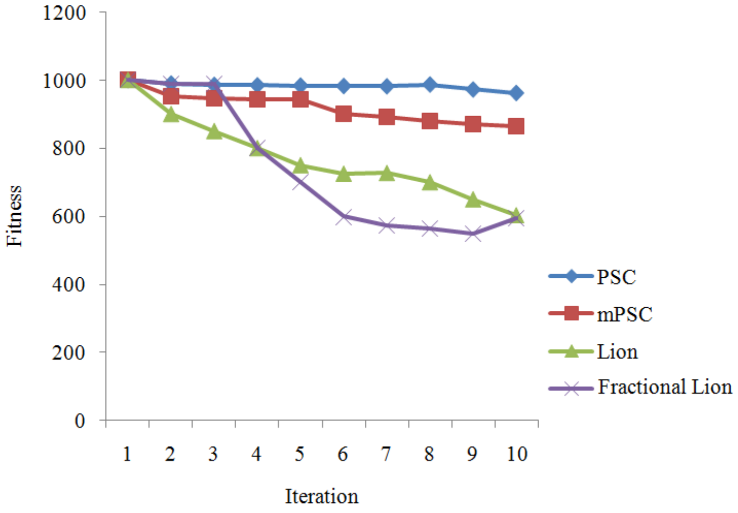


Fig. 8. Convergence analysis of Wine datasets cluster 3

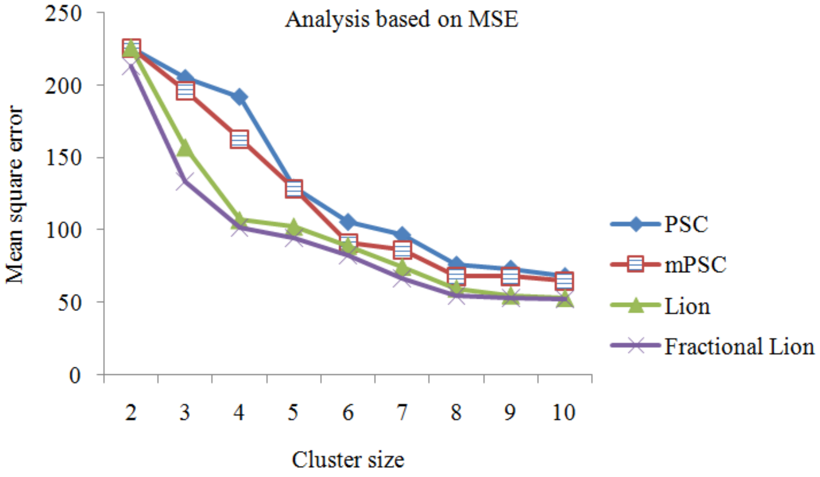


Fig. 9. Analysis based on MSE for iris data set

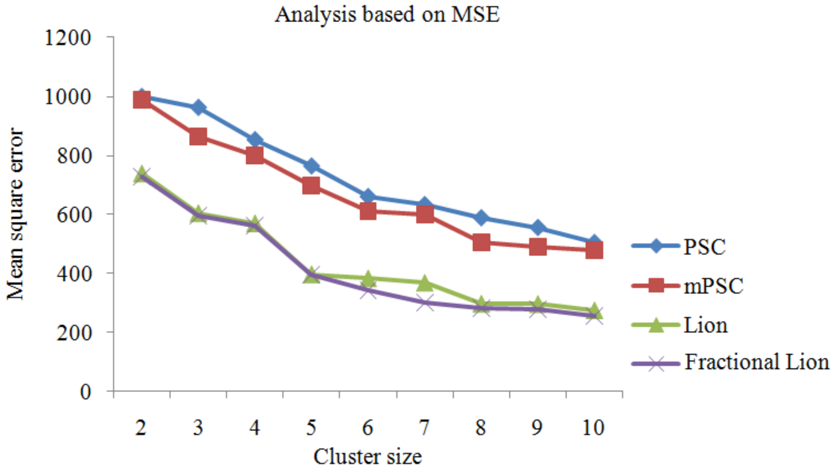


Fig. 10. Analysis based on MSE for wine data set

Despite of repetitive iteration in both the clusters, the fitness value is always lower than the existing algorithm thereby making the proposal an efficient one for clustering.

Analysis Based on Mean Square Error

The experimental results for the benchmarked data sets based on the mean square error are viewed in this section. The MSE analyses over the different cluster size are analysed.

Figure 9, shows the performance curve for the MSE analysis of iris data sets. In iris dataset, for the cluster of size 2, the means square error value obtained by the proposed algorithm is 213. But at the cluster of size 2, the MSE value for the PSC, mPSC and lion algorithm are 225.56, 225.56 and 225.56 respectively. On comparing the values, the proposed estimation algorithm error value is small over the others. The minimum error value enables the clustering with increased speed and ease in searching the centroid point. The increase in the cluster size decreases the mean square error value. At the 10th cluster, the MSE value obtained for the fractional lion is 52.03.

Figure 10, shows the MSE analysis curve of wine data set. At the 2 cluster, the MSE value obtained from the existing algorithms are 1000, 987.16, 737.9 respectively. The value out of the proposed algorithm is 726.95. The difference in the MSE value calculated is about 250 for the PSC algorithm and around 260 for the mPSC algorithm respectively at the 2nd cluster. The MSE value of 256.37 is obtained for the cluster size of 10 which is very much lower considering the existing algorithm. Thus the minimum value in the error enables the proposed algorithm a well suited one for the rapid centroid estimation for the clustering.

Analysis Based on Rand Coefficient

Figure 11 and 12, shows the performance evaluation curve for the fractional lion algorithm based on the rand coefficient. The performance curve is plotted between the cluster size and rand coefficient.

Figure 11, represent the performance curve plotted between the cluster size and rand coefficient for the iris data set. The rand values must be minimum for the improved clustering performances. From the curve, the value obtained for the rand coefficients for the changeover cluster size from 2 to 10, is between 77.49 and 78.45. The increment in the value is about 1 for the fractional lion. But the change in the rand coefficient obtained at the PSC, mPSC and lion algorithm are of the range 5, 1.3 and 3 respectively which is higher comparing the proposed algorithm.

The performance curve for the wine data set based on the rand coefficient is shown in Fig. 12. The cluster size of 2,3,4,5 is used for the random coefficient validity measures. At the cluster size 2, the random coefficient value obtained is 61.36, 69.98, 75.79 and 76.24 respectively for the PSC, mPSC, Lion and fractional Lion algorithm. The value of rand indices for the proposed algorithm is increased by about 15 over the existing algorithm. The rand coefficient values for the cluster sizes 3, 4, 5 are around 77 which are higher than the values out of all the existing algorithm.

Analysis Based on Jaccard Coefficient

The performance curve analysing the proposed algorithm concerned with the jaccard coefficient are shown in Fig. 13. The jaccard coefficient is one of the clustering validity measures. The coefficient shows the similarity between the proposed and exiting clustering output.

For iris data set, the performance curve is plotted between the cluster size and jaccard coefficient in Fig. 14. The experimentation is repeated from the cluster size 2 to the cluster size 10. From the curve, it is clear that the obtained output have the decreased jaccard values over the cluster sizes. The jaccard value of 35.81 is obtained at the 8th cluster whereas the jaccard indices for the exiting algorithms are 44.43, 43.81 and 48.75 respectively. The value shows a fractional increase in the validity coefficient over the proposed algorithm.

Figure 14, shows the performance curve plotted between the cluster size and jaccard index. The experimentation is repeated for cluster size about 2, 3, 4 and 5. For the cluster size 4, the jaccard coefficient values obtained for the PSC, mPSC, Lion and fractional Lion are 56.39, 61.58, 62.34 and 63.48 respectively. The value out of the proposed experimentation is higher compared to the existing clustering algorithm.

The analysis based on the rand and jaccard coefficient confirms that the proposed fractional lion algorithm is suitable for the clustering with multi cluster data sets and clusters with different scales.

Analysis Based on Clustering Accuracy

The performance evaluation based on the clustering accuracy is described in this section. Figure 15 and 16 represent the performance curve based on the clustering accuracy for the iris and wine data sets.

Figure 15, shows the performance curve relating the clustering accuracy of the proposed algorithm in the iris data set. The clustering accuracy attained by the fractional lion algorithm is similar to that of the accuracy resulting from the lion algorithm. For the cluster size 3, the clustering accuracy achieved by the PSC and mPSC are 69.3 and 70% whereas the proposed algorithm have the accuracy of 75%. The accuracy achieved is about 5% higher than that of the PSC and mPSC. The overall accuracy achieved is approximately 8% higher compared to the existing clustering algorithm.

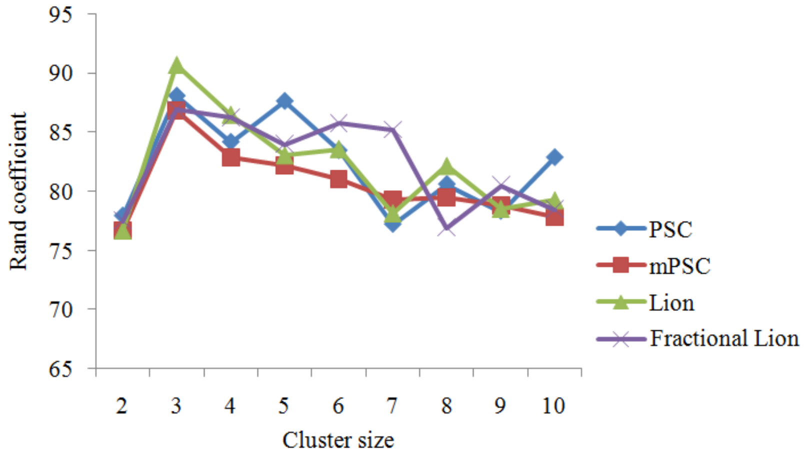


Fig. 11. Analysis based on rand coefficient for iris dataset

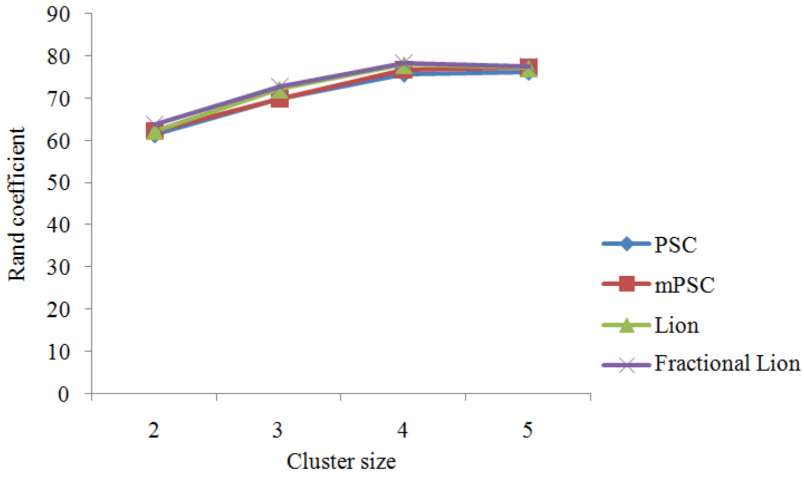


Fig. 12. Analysis based on rand coefficient for wine dataset

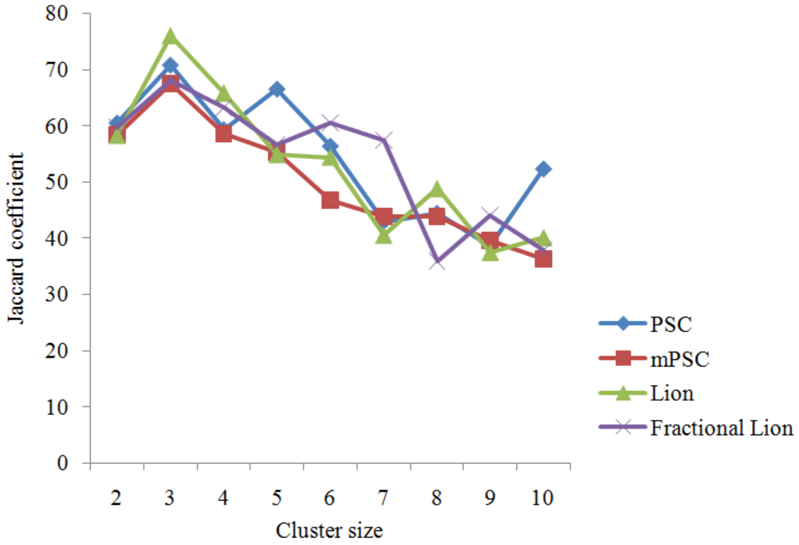


Fig. 13. Analysis based on Jaccard coefficient foriris dataset

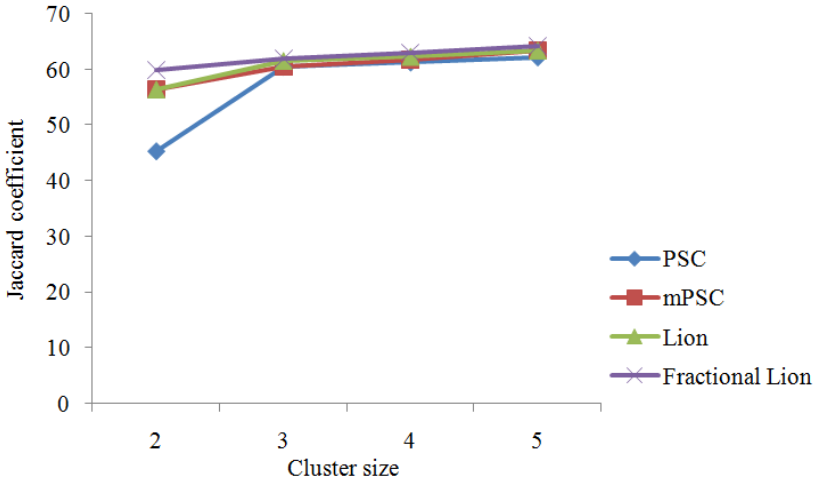


Fig. 14. Analysis based on Jaccard coefficient for wine dataset

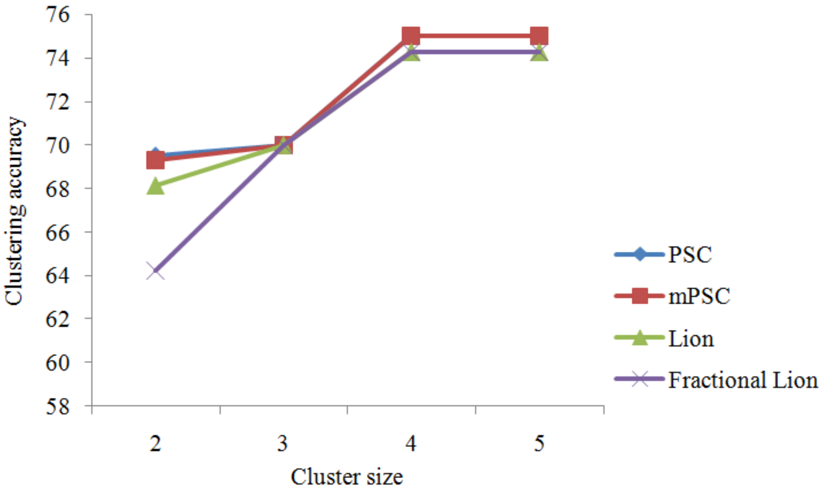


Fig. 15. Analysis based on clustering accuracy for iris dataset

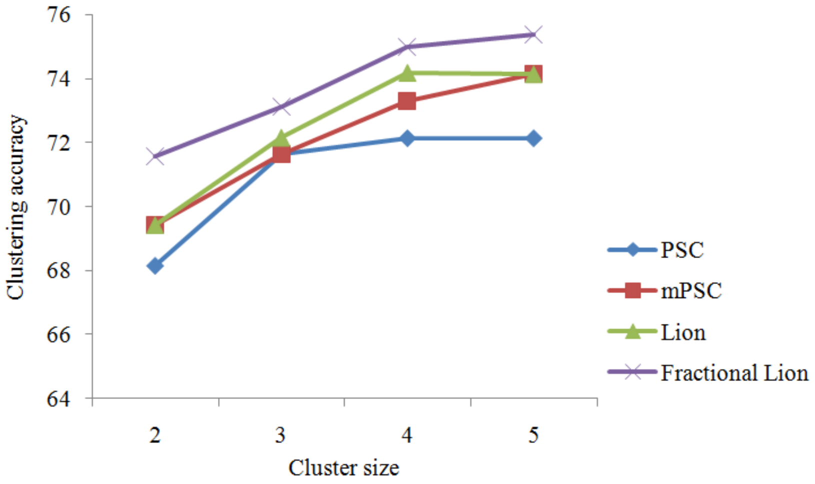


Fig. 16. Analysis based on clustering accuracy for wine dataset

The performance curve analysing the clustering accuracy of the wine data set is shown in the Fig. 16. For the wine data set, the accuracy is improved about 1% compared to the lion algorithm. Thus proposed algorithm proves to be more effective one for the clustering procedure. At the cluster size 4, the accuracy attained by fractional lion is 74.15% which is 4.73% higher than the PSC algorithm, 1.99% higher than the mPSC algorithm and 0.05% higher than the lion algorithm.

From the obtained value, the proposed algorithm is evident to be effective clustering algorithm with improved clustering accuracy.

Discussion

The proposed FLA for clustering is validated with the external evaluation metrics like, rand coefficient, jaccard coefficient and clustering accuracy for showing the performance improvement. While discussing with the other work, the proposed shows the rapid estimation of cluster centroids. Accordingly, Huang *et al.* (2014) have proposed clustering algorithm by expanding the existing k-means-type algorithms which does not include nay rapid measurement mechanism for the speed up the clustering procedure. Also, Parker and Hall (2014) have proposed two accelerated algorithms, such as Geometric Progressive Fuzzy C-Means (GOFCM) and Minimum Sample Estimate Random Fuzzy C-Means (MSERFCM) which apply statistical method to calculate the subsample size. Even though they considered fuzzy theory for clustering, the estimation of sample size do not provide the accurate clustering output. Ji *et al.* (2012) have combined mean and fuzzy centroid to symbolize the prototype of a cluster and used co-occurrence of values based measure to examine the difference between data objects and prototypes of clusters. This algorithm lack of finding the minimu distance measurement among the data points which is included n the proposed FLA clustering. Yin *et al.* (2010) has proposed an adaptive Semi-supervised Clustering Kernel Method depending on Metric learning (SCKMM) but the inclusion of optimization algorithm is completely missed in this study so the rapid estimation seems challenging one.

Accordingly, Yuwono *et al.* (2014) have proposed an algorithm called, PSC for solving clustering problems. PSC algorithm for substitution was constructed such that it is easy and efficient to implement but the speed of operation to convergence seems tough. Sheng *et al.* (2014) have used three local searches of various features to effectively utilize the decision space but the combination of three search methods require more computational time to reach the optimal value. Bandyopadhyay (2011) have presented a combination of a measure for cluster validity and the concept of stability to describe two objective functions which are concurrently optimized for clustering. This method finds lack of rapid estimation due to the multiple objective constraints. Kiranyaz *et al.* (2010) have proposed two methods that effectively concentrate on various severe issues in the area of Particle Swarm Optimization (PSO). Kuo *et al.* (2012) have proposed a dynamic clustering technique depending on Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) (DCPG) algorithm. Commonly, PSO and GA algorithms do not fit for the complex searching tasks like clustering and classification. The proposed FLA finds the advantage of fast convergence and rapid centroid estimation as compared with the existing algorithms.

More commonly, clustering finds a lot of application in various fields, such as telecommunication, medical data and electrical applications and so on. More specifically, diagnostic system for disease, trading system, optimizing social regulation policies, self-integrating knowledge-based system, integrating design stages, University admission process, genetic mining, topic based on concept distribution, Intelligent web miner, neural network, extraction of fuzzy classification rules, intrusion detection system, text mining, search engine, signal processing, image processing are some example of applications which mostly utilized effective clustering process.

Conclusion

This paper presented a novel fractional lion algorithm for the rapid centroid estimation in the data clustering. Here, FLA was proposed by combining the modified fractional theory to the Lion algorithm which is based on the lion pride behaviour. The fitness evaluation was also developed for the selection of the optimal centroid point from the solution points. The proposed FLA algorithm was successfully adapted with the clustering procedure. Since the iteration is continued till the tolerance is achievable only the best solution point is chosen as the optimal centroid. The experimentation is performed using benchmarked iris and wine datasets. The performance of the proposed system is compared with the Particle Swarm Clustering algorithm (PSC), modified Particle Swarm Clustering algorithm (mPSC) and lion algorithm. The performance is evaluated using the clustering accuracy, rand and jaccard coefficient and MSE metrics. The proposed FLA obtained rapid estimation of centroid even in bulk data sets. The result obtained out of the proposed algorithm is efficient with improved clustering accuracy than that of the existing clustering algorithm. In future, multi kernel-based distance measurement can be included for finding the fitness of the cluster process. Also, the neighbour solution defined within FLA algorithm can be further strengthened with different optimization theory.

Funding Information

The authors have no support or funding to report.

Author's Contributions

Satish Chander: Conception, Design, Implementation and Analysis of data.

P. Vijaya: Conception, Design and Writing the manuscript.

Praveen Dhyani: Scientific advisor for the overall coordination and final approval of the research work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Bandyopadhyay, S., 2011. Multiobjective simulated annealing for fuzzy clustering with stability and validity. IEEE Trans. Syst. Man cybernet. Part C: Applic. Rev., 41: 682-691. DOI: 10.1109/TSMCC.2010.2088390

Binu, D., 2015. Cluster analysis using optimization algorithms with newly designed objective functions. Expert Syst. Applic., 42: 5848-5859. DOI: 10.1016/j.eswa.2015.03.031

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybernet., 3: 32-57. DOI: 10.1080/01969727308546046

Filho, T.M.S., B.A. Pimentel, R.M.C.R. Souza and A.L.I. Oliveira, 2015. Hybrid methods for fuzzy clustering based on fuzzy c-means and improved particle swarm optimization. Expert Syst. Applic., 42: 6315-6328. DOI: 10.1016/j.eswa.2015.04.032

Huang, X., Y. Ye and H. Zhang, 2014. Extensions of kmeans-type algorithms: A new clustering framework by integrating intracluster compactness and intercluster separation. IEEE Trans. Neural Netw. Learn. Syst., 25: 1433-1446. DOI: 10.1109/TNNLS.2013.2293795

Ji, J., W. Pang, C. Zhou, X. Han and Z. Wang, 2012. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Know. -Based Syst., 30: 129-135. DOI: 10.1016/j.knosys.2012.01.006

Kiranyaz, S., T.A. Yildirim and M. Gabbouj, 2010. Fractional particle swarm optimization in multidimensional search space. IEEE Trans. Syst. Man Cybernet. Part B: Cybernet., 40: 298-319.

Kotsiantis, S. and P. Pintelas, 2004. Recent advances in clustering: A brief survey. WSEAS Trans. Inform. Sci. Applic., 1: 73-81.

Kuo, R.J., Y.J. Syu, Z.Y. Chen and F.C. Tien, 2012. Integration of particle swarm optimization and genetic algorithm for dynamic clustering. Inform. Sci., 195: 124-140. DOI: 10.1016/j.ins.2012.01.021

Lee, J.S. and S. Olafsson, 2005. Data clustering by minimizing disconnectivity. Inform. Sci., 181: 732-746. DOI: 10.1016/j.ins.2010.10.028

Lichman, M., 2013. UCI machine learning repository. University of California, Irvine, CA.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, (MSP' 67), Berkeley, Calif., pp: 281-297.

Maji, P., 2011. Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data. IEEE Trans. Syst., Man Cybernet. Part b: Cybernet., 41: 222-253. DOI: 10.1109/TSMCB.2010.2050684

Parker, J.K. and L.O. Hall, 2014. Accelerating fuzzy-c means using an estimated subsample size. IEEE trans. Fuzzy Syst., 22: 1229-1244. DOI: 10.1109/TFUZZ.2013.2286993

Pires, E.J.S., J.A.T. Machado, P.B. de Moura Oliveira, J.B. Cunha and L. Mendes, 2010. Particle swarm optimization with fractional-order velocity. Nonlinear Dynam., 61: 295-301. DOI: 10.1007/s11071-009-9649-y

Rajakumar, B., 2012. The lion's algorithm: A new nature-inspired search algorithm. Procedia Technol., 6: 126-135. DOI: 10.1016/j.protcy.2012.10.016

Sheng, W., S. Chen, M. Fairhurst, G. Xiao and J. Mao, 2014. Multilocal search and adaptive niching based memetic algorithm with a consensus criterion for data clustering. IEEE Trans. Evolutionary Comput., 18: 721-741. DOI: 10.1109/TEVC.2013.2283513

Wan, M., L. Li, J. Xiao, C. Wang and Y. Yang, 2012. Data clustering using bacterial foraging optimization. J. Intell. Inf. Syst., 38: 321-341. DOI: 10.1007/s10844-011-0158-3

Xu, R. and D. Wunsch, 2005. Survey of clustering algorithms. IEEE Trans. Neural Netw., 16: 645-677.

Yin, X., S. Chen, E. Hu and D. Zhang, 2010. Semi-supervised clustering with metric learning: An adaptive kernel method. Patt. Recognit., 43: 1320-1333. DOI: 10.1016/j.patcog.2009.11.005

Yuwono, M., S.W. Su, B.D. Moulton and H.T. Nguyen, 2012. Method for increasing the computation speed of an unsupervised learning approach for data clustering. Proceedings of the IEEE CEC, (CEC12), pp: 2957-2964. DOI: 10.1109/cec.2012.6252927

Yuwono, M., S.W. Su, B.D. Moulton and H.T. Nguyen, 2014. Data clustering using variants of rapid centroid estimation. IEEE Trans. Evolut. Comput., 18: 366-377. DOI: 10.1109/TEVC.2013.2281545