# FGPSO - A Novel Algorithm for Multi Objective Data Clustering

APARNA K[1], MYDHILI K NAIR[2]
[1]Department of Master of Computer Applications
BMS Institute of Technology & Management
Avalahalli, Yelahanka, Bengaluru – 560 064, Karnataka State, INDIA

[2]Department of Information Science & Engineering
M S Ramaiah Institute of Technology
MSR Nagar, Mathikere, Bengaluru – 560 054, Karnataka State, INDIA
[1]aparnak.bmsit@gmail.com , [2]mydhili.nair@gmail.com

*Abstract:* - The task of clustering is to group the data items that are similar into different clusters in such a way that the similarity within each cluster is high and the dissimilarity between the clusters is also high. A novel partitional clustering algorithm called HB K-Means algorithm (High Dimensional Bisecting K-Means) based on high dimensional data set was developed in our previous work. In order to improve this novel algorithm, constraints such as Stability based measure and Mean Square Error (MSE) were incorporated resulting in CHB K-Means (Constraint Based HB K-Means) algorithm. In addition to these constraints, cluster compactness and density are also important to obtain better clustering results. In this paper, it is proposed to develop a Multi-Objective Optimization (MOO) technique by including different indices such as DB-Index, XB-Index and Sym-Index. These three indices will be used as fitness function for the proposed Fractional Genetic PSO algorithm (FGPSO) which is the hybrid optimization algorithm to do the clustering process. The performance of this optimization algorithm is evaluated based on parameters such as Clustering Accuracy and Time Computation by executing the algorithm on some of the benchmark datasets taken from UCI Machine Learning Repository.

*Key-Words:* - Partitional Clustering, Multi-Objective Optimization, DB Index, XB-Index, Sym-index, Fractional Genetic PSO Algorithm (FGPSO).

## 1 Introduction

Knowledge Discovery in Databases (KDD) consists of many tasks in order to process the raw data into useful information including Data Mining. Clustering, also known as Unsupervised Classification is one of the major tasks of Data Mining which involves grouping of similar data objects into clusters. A detailed study and analysis of the different partitional clustering algorithms is given in [1]. A wide range of real-time problems involves clustering algorithms along with exploratory data analysis [4], image segmentation [5] and mathematical programming [6, 7]. In order to arrive at optimized solution, clustering techniques have also been applied effectively to deal with the scalability problem of machine learning [8, 9, and 10]. Clustering helps in analysing the collection of input data in order to derive out useful patterns [11]. These patterns can be represented in terms of mathematical vector in the multi-dimensional space.

The two main categories of clustering algorithms are classified as supervised and unsupervised clustering. The lack of class information differentiates unsupervised learning (clustering) supervised learning (classification). It is not possible to access any labelled data in unsupervised classification [12, 13]. Clustering has the objective of dividing an unlabeled data set into a fixed and separate set of useful data patterns [14, 15].

Many clustering algorithms are proposed till date. One of the most widely used clustering algorithms is the K-Means algorithm which divides the data objects into K clusters [16]. The data objects can be allocated to multiple clusters using Fuzzy algorithms. One of the efficient algorithms in this category is the Fuzzy C-Means algorithm. Moreover the arbitrary choice in initializing the centre points makes the iterative process in achieving local optimal solution without difficulty. Many evolutionary algorithms are available in order to improve such solutions, such as Genetic Algorithm (GA) [17], Simulated Annealing (SA) [18], Ant Colony Optimization (ACO) [19], and Particle Swarm Optimization (PSO) [20]. Multi-objective clustering can be looked out as a unique case of multi-objective optimization which plans to

concurrently optimize multiple objectives under definite constraints.

The rest of the paper is organized as follows: the recent research works is analyzed in section 2; the proposed work is briefly explained in section 3; in the section 4, the experimental results along with the comparison analysis are depicted and the section 5 represents the conclusion of the paper.

## 2 Literature Survey

The literature review reveals a lot of study done based on Multi Objective data clustering. Some of them are presented here: The authors in [21] have proposed that the dimensions of a high dimensional dataset can be reduced more efficiently and effectively using Canonical Variate analysis. The modified K-Means algorithm is then applied to the reduced low dimensional dataset. In order to further optimize the solution, Genetic algorithm is applied for the purpose of initializing the centroids of the Improved Hybridized K-Means Clustering Algorithm (IHKMCA) in their paper. As compared to other approaches, the work has shown effective and accurate results with less time consumption.

Tulin Inkaya *et al* [22] have described a novel methodology called Ant Colony Optimization based Clustering methodology (ACO-C). A few limitations of the clustering problem including solution assessment, neighbourhood construction and data set reduction are tackled by this methodology. Two objective functions namely, adjusted compactness and relative separation are used in this framework in order to assess the clustering solution based on the local features of the neighbourhoods. This helps to measure the quality of clustering solutions without using prior data. Two pre-processing steps are involved in ACO-C, namely, neighbourhood construction and data set reduction. The local features of the data points are removed using neighbourhood construction and the stability is achieved using data set reduction.

In [23] the authors have proposed an interval weighted fuzzy C-Means clustering by genetically guided alternating optimization technique where the interval number was considered for attribute weighing. The authors have demonstrated that from the point of view of geometric probability, the attained interval weighing was suitable. Moreover, a genetic heuristic approach for attribute weight searching was also considered to direct the alternating optimization (AO) of WFCM. The experimental results showed that the algorithm performed better. It exposes the interval weighted clustering as an optimization operator on the basis of the traditional numerical weighted clustering, and the results of the interval weight perturbation on clustering performance was reduced.

The authors in [2] have brought out a novel partitional clustering algorithm called HBK-Means Algorithm. In this algorithm, the high dimensional dataset is converted to Attribute Frequency Matrix and it is then clustered using the modified Bisecting K-Means algorithm. The experimentation is carried out on two large datasets of UCI machine learning repository and the proposed algorithm has achieved better clustering accuracy and smaller computation time compared to the traditional K-Means clustering algorithm.

In [3], the authors have incorporated two constraints such as, Stability-based measure and Mean Square Error (MSE) on the novel partitional clustering method, known as CHB-K-Means (Constraint based High dimensional Bisecting K-Means) algorithm to improve the performance of HBK-Means algorithm [2]. The CHB-K-Means algorithm generates two initial partitions. Subsequently, it calculates the Stability and MSE for each partition generated. Inference techniques are applied on the Stability and MSE values of the two partitions to select the next partition for re-clustering process. This process is repeated until K number of clusters is obtained. From the experimental analysis, it can be inferred that an average clustering accuracy of 75% has been achieved. The comparative analysis of the proposed approach with the other traditional algorithms shows not only an achievement of higher clustering accuracy rate but also a decline in utilization of execution time.

## 3 Proposed Multi Objective Fractional Genetic PSO (FGPSO) Algorithm for Data Clustering

The most serious task associated with data mining is the assessment of the number of clusters required to execute the clustering algorithm. This difficult issue can be solved by using an optimization technique that involves a multi-objective function. In the past, the effectiveness of clustering was computed with several kinds of objective functions. But, currently, improved accuracy can be produced with the use of multi-objective optimization technique. This paper concentrates more on the development of Multi-Objective Optimization (MOO) technique that involves various indices such as, DB index, sym-index and XB-index. These three indices will constitute the fitness function, which helps in

performing the proposed Fractional Genetic PSO algorithm (FGPSO) in an efficient way. This algorithm is a hybrid optimization algorithm that carries out the clustering process. Fig. 1 portrays the basic block architecture of the Fractional Genetic PSO algorithm.

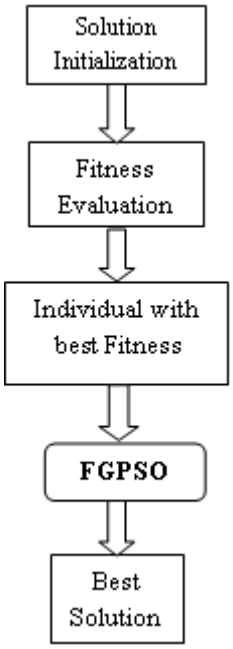


Fig.1 Block Diagram of the proposed FGPSO

As indicated in Fig.1, the proposed multi-objective data clustering has solution initialization as its primary step. Once the solutions have been initialized, the objective functions involving DB index, sym-index and XB-index are considered for evaluating the fitness of each solution. At the end, clustering process is carried out with the proposed FGPSO algorithm. In the population initialization phase of FGPSO, which forms the initial phase of the algorithm, the generation of population is achieved. Here, both the input data as well as the generated initial solution are found to have the same size.

## 3.1 Solution Initialization

In the proposed multi-objective data clustering approach, an arbitrary generation of initial solution is done first. The input data and the initialized solution should be of identical size. The length of each and every single individual solution belonging to the population is *n* and then, each data point of the solution will be allocated to the K clusters in a random fashion. At the initial stage, current generation is assumed a value of zero. All the chromosomes will have a number of cluster parameters that lie between one and the user-specified maximum number of clusters. A population containing a certain number of chromosomes is generated in an arbitrary way. In the beginning, random assignment of data points to every single cluster takes place. Next, the remaining points are also arbitrarily allocated to the clusters. This method prevents the illegal strings to be generated. Illegal strings are those strings with no pattern in any of the cluster.

The proposed FGPSO algorithm has the fitness evaluation as its subsequent step. Three different indices, namely, DB index, XB-index and sym-index allow the fitness of each solution to be computed.

## 3.2 Fitness Calculation

The measure of an individual's quality is provided using the fitness function. Designing of fitness function is essential to evaluate the individual's performance in the present population. While information retrieval has to be performed in the genetic algorithm, it is necessary to provide an evaluation or fitness function for all the problems to be solved. The choice of more appropriate fitness function is important because the operation of the genetic algorithm relies on it. The fitness function in the proposed approach aids in finding the proper feature subsets. The various fitness functions are explained as follows:

### 3.2.1 DB-Index

The Davies Bouldin Index allows the clustering algorithm to be assessed. The DB-Index is an internal evaluation scheme, which validates the manner in which the clustering is accurately performed in accordance with the quantities as well as the features related to the dataset. The computation of the quality of clustering can be enhanced with the inclusion of the Fuzzy matrix to the Davies Bouldin index. The computation of Fuzzy DB-Index is elucidated below.

$$DBI = \frac{1}{c}\sum_{l=1}^{c} R_{a,a+1} \qquad (2)$$

where, *DBI* is the Fuzzy Davies Bouldin Index, $R_{a,\,a+1}$is the clustering scheme measurement between each cluster and *c* is the total number of clusters. The clustering scheme measurement between each cluster, $R_{a,\,a+1}$, can be computed with the following equation.

$$R_{a,a+1} = \frac{S_a + S_{a+1}}{M} \quad (3)$$

In the aforementioned equation, $S_a$ specifies the distance value between each data in the cluster and the centroid associated with that cluster; $S_{a+1}$ indicates the distance value between each data in the subsequent cluster and the centroid of that cluster; $M$ refers to the sum of the Euclidean distance between each centroid and the value of $M$ can be computed using the equation stated below.

$$M = \sum_{a=1}^{c-1} \sum_{d=a+1}^{c} \sqrt{(C_a - C_d)^2} \quad (4)$$

The measurement of the distance value between each data in the cluster and its centroid involves the following equation.

$$S_a = \frac{1}{T} \sum_{b=1}^{T} \left| X_b - C_a \right|^2 \times U_b \quad (5)$$

where $T$ is the total number of data in the set, $X$ is the data in the cluster and $C$ is the centroid of the cluster. Here, $U_b$ represents the computation of Fuzzy matrix.

### 3.2.2 Sym-Index

Most appropriate clustering results can be produced, when a maximum value is obtained for the Sym-Index. The Sym-Index can be expressed as:

$$SI = \frac{1}{c} \times \frac{1}{E} \times g \quad (6)$$

where $c$ is the total number of clusters, $E$ is the sum of the Euclidean distance between data and centroids in all clusters, and $g$ is maximum distance between centroids. The sum of the Euclidean distance between the centroids and its corresponding data can be evaluated using the following equation.

$$E = \sum_{a=1}^{c} \sum_{b=1}^{T} \left| C_a - X_b \right|^2 \quad (7)$$

In the above equation, $T$ specifies the number of data that is related to the centroid $C_a$. For example, if $C_1$ indicate the centroid of the first cluster, then $X_b$ are the data in the first cluster. The value of $g$ is opted from the comparison of distance between the centroids that rely on the maximum value and it can be provided with the equation below:

$$g = \max\{dis(C_a, C_d)\} \quad (8)$$

In the above equation, $a$ falls in the range of $1$ and $c - 1$; $d$ lies between $a+1$ and $c$. The above equation makes a comparison of the distance between each centroid and selects the maximum distance value.

### 3.2.3 XB-Index

The XB-Index can be defined as the ratio of the within cluster compactness to the minimum distance between the clusters and it is given by,

$$XBI = \frac{E}{m \times h} \quad (9)$$

In the above equation, $E$ points to the sum of the Euclidean distance between the centroids and its related data; m specifies the total number of data in the dataset of very high dimension; the value of $h$ is provided using the comparison of distance between the centroids in accordance with the minimum value and it is given as follows:

$$h = \min\{dis(C_a, C_d)\} \quad (10)$$

In the above equation, $a$ takes a value between $1$ and $c - 1$; $d$ assumes value from $a + 1$ to $c$. The above equation makes a comparison of the distance between each centroid and it would opt for the minimum distance value. The computation of Fuzzy DB-Index, Sym-Index and XB-Index allows the fitness value computation of each nest to be achieved.

## 3.3 Proposed Fractional GPSO

The proposed fractional GPSO algorithm starts with the random initialization of the solutions. Next, the fitness function involving the three different indices is assessed for every single particle. For each and every iteration, the Particle best ($P_{best}$) and global best ($g_{best}$) among the initialized solution should be computed. In addition, a group of parent solutions are chosen to carry out the genetic operation. The set of parent solutions, which have been chosen from the population are the global best solution ($g_{best}$) and the worst solution. A fresh solution is achievable, if crossover and mutation operation are applied on the chosen set of parent solution.

### 3.3.1 Crossover Operation

Assume that there are two parents. Let P = {p1, p2, p3, p4} and Q = {q1, q2, q3, q4} indicate the parent solutions containing four cluster centers, where each pi and qi represents a vector of features as shown in the fig.2 below. Two children A and B are produced as a result of crossover. If the uniform crossover is considered, the crossover on every single pair of centers from the parents can be determined using a probability of 0.5. In the following example, the

crossover operation is neglected at position 3 and hence, the values of p3 and q3 are copied to position 3 of strings A and B in order and in a straightforward manner. Positions 1, 2 and 4 take new values of a1, a2, a4 and b1, b2, b4 on A and B, respectively.

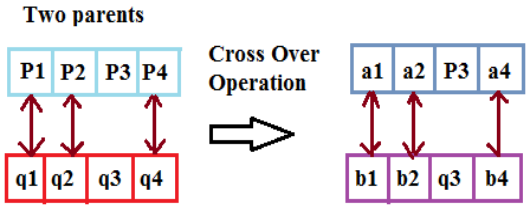


Fig. 2 Crossover operation achieved on the chosen parent

### 3.3.2 Mutation Operation

Random mutation is utilized in this paper and hence, only a small fraction of the bits contained in the chromosome list gets modified. Mutation serves as an alternative approach to find the cost surface. It is capable of introducing new attributes, which the original population lacks, and helps the genetic algorithm to have a rapid convergence. A new solution will be obtained only after the application of crossover and mutation on the parent solution set.

### 3.3.3 FPSO Operation

This operation makes use of the FPSO operators that include the velocity updates as well as the position updates for revising the solutions containing the worst fitness. The velocity and the position of the particles in the FPSO algorithm can be revised using the following expression. The velocity updating equation of FPSO is stated as follows:

$$vel_{t+1} - vel_t = \frac{1}{2}\alpha vel_{t-1} + \frac{1}{6}\alpha(1-\alpha)vel_{t-2} + \frac{1}{24}\alpha(1-\alpha)(2-\alpha)vel_{t-3} + \psi_1(p_{best} - pos) + \psi_2(g_{best} - pos) \quad (9)$$

where α indicates a small positive constant. Likewise, the position updating equation in the FPSO algorithm is given by,

$$pos_{t+1} = pos_t + vel_{t+1} \quad (10)$$

Immediately after the updation of the position as well as the velocity of the particle, the fitness for the new solution is computed. The solution containing worst fitness is chosen and once more the genetic operator will be applied. The new solutions will then have the updated velocity and position. The process will be executed continuously until the ceasing condition has been met. The various steps involved in the proposed FPSO algorithm are portrayed below.

### 3.3.3 Step by Step Procedure of FGPSO Algorithm

**1. Initialization.** Generate initial population randomly with size equal to the input data.
**2. Fitness calculation**. Evaluate the fitness for each solution in the population based on five objective functions.
**3. GA operation**. A set of parent solution is selected and the crossover and mutation operation is performed to generate new solution.
**4. FPSO operation**. The solution with worst fitness is updated using the velocity and position updating operation
The velocity updating equation is

$$vel_{t+1} - vel_t = \frac{1}{2}\alpha vel_{t-1} + \frac{1}{6}\alpha(1-\alpha)vel_{t-2} + \frac{1}{24}\alpha(1-\alpha)(2-\alpha)vel_{t-3} + \psi_1(p_{best} - pos) + \psi_2(g_{best} - pos)$$

Also the position updating equation is given below,

$$pos_{t+1} = pos_t + vel_{t+1}$$

**5.** Calculate the fitness for the new solution
**6.** Apply GA operation between new solution and old solution at K iteration
**7.** Update new best solution using FPSO operator
**8.** Repeat the steps 2 to 4 until the termination criteria reached.

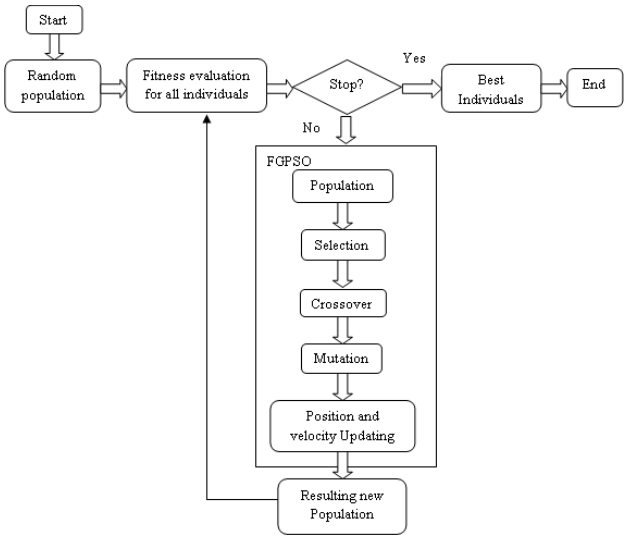


Fig.3. Flow Diagram of the proposed FGPSO

The flow diagram of the proposed Fractional Genetic PSO algorithm for multi objective data Clustering is shown in the fig. 3 above.

# 4 Results and Discussion

The proposed multi objective data clustering is implemented in the MATLAB and the data clustering is experimented with the dataset. The datasets used to compare the performance of the proposed approach are taken from UCI machine learning repository which includes the Spam base Dataset, Localization Data for Person Activity Data Set, Pen-based recognition of handwritten digits dataset. The suggested multi objective data clustering is executed in a Windows machine containing configurations Intel (R) Core i5 processor, 1.6 GHz, 4 GB RAM, and the Operating System platform is Microsoft Windows 7 Professional. Three datasets are used here namely Spambase Dataset, Localization Data for Person Activity Data Set, and Pen-based recognition of handwritten digits dataset which are taken from the UCI machine learning repository. The detailed explanation of dataset is given below.

## 4.1 Quantitative Measure

The evaluation of the proposed multi objective data clustering is carried out using the following metrics.

### 4.1.1 Clustering Accuracy

Clustering accuracy refers to the degree of closeness of measurement of a quantity to its actual value. In this paper the clustering accuracy is computed using the following formula.

$$CA = \frac{1}{N}\sum_{i=1}^{T} X_i \qquad (11)$$

where, $N$ is the number of data point and $T$ is the number of class and X is the concerned data point.

### 4.1.2 Computation Time

Computation time or time complexity evaluates the measure of time taken by an algorithm to execute. Likewise, it decides how the execution time varies with size and dimensionality of the dataset.

## 4.2 Performance Evaluation

The basic idea of the proposed approach is multi objective data clustering using Fractional Genetic PSO (FGPSO) algorithm. To prove the efficiency of our method, we compare the proposed FGPSO algorithm to K-Means algorithm, Bisecting K-Means algorithm, HB K-Means [2], MSE-HB-K-Means and Stable-HB-K-Means (both together is referred to as CHB K-Means) [3], Genetic algorithm and Particle Swarm Optimization algorithm. In this section we evaluate our proposed method based on clustering accuracy and computation time. All the analyses were done on the Spambase dataset [24], Localization Data for Person Activity dataset [25] and the Pen-Based Recognition of Handwritten Digits Dataset [26].

### 4.2.1 Performance Analysis Based on Clustering Accuracy

In this section, we plot the performance analysis of the proposed method, Fractional Genetic PSO (FGPSO) algorithm. The performance of the proposed approach is evaluated based on the clustering accuracy. The process is conducted by varying the number of clusters.

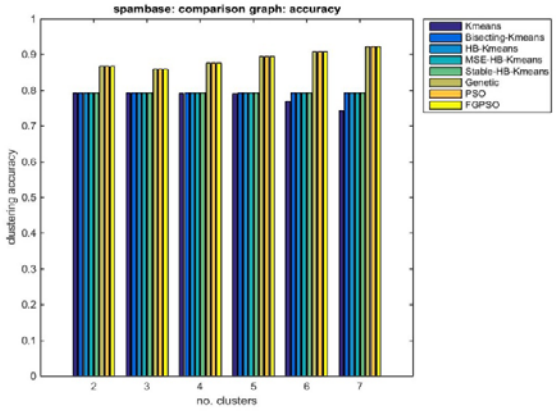


Fig.4. Performance based on clustering accuracy using dataset 1

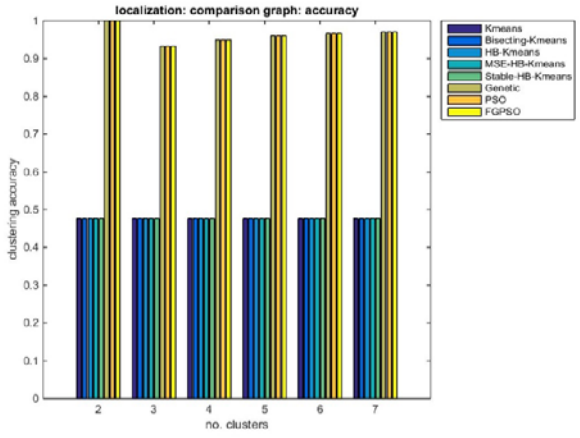


Fig.5. Performance based on clustering accuracy using dataset 2

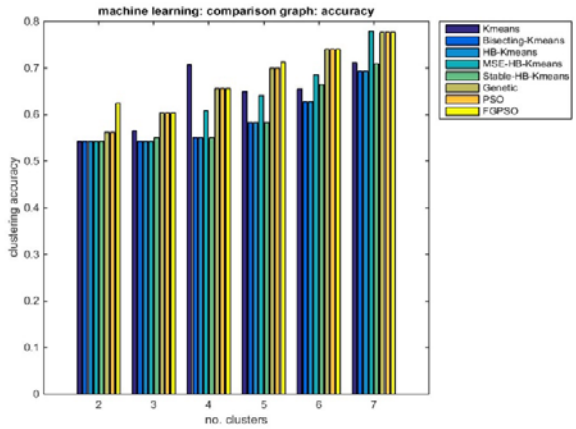


Fig.6. Performance based on clustering accuracy using dataset 3

From the result shown in fig. 4, one can observe that PSO, GA and our proposed method FGPSO yield the best performances followed by K-Means algorithm, Bisecting K-Means algorithm, HB K-Means, MSE-HB-K-Means and Stable-HB-K-Means. But our proposed approach which is implemented using GA and PSO is slightly better than the existing approaches. When the cluster value is 2 we obtain the maximum accuracy value using GA, PSO and FGPSO. The fig. 5 illustrates the performance of the proposed approach using dataset 2. Here, when the cluster value is 2, we obtain the accuracy value of 65%, which is very much high compared to all the existing approaches. Fig. 6 shows the performance comparison of accuracy plot using dataset 3. Here, when number of clusters is increased to 7, we achieve the maximum accuracy value for GA, PSO and FGPSO approaches. We can conclude from the responses of dataset 1 and dataset 2 that as the number of clusters increases, the clustering accuracy also increases accordingly.

### 4.2.2 Performance Analysis Based on Computation Time

In this section, we discuss performance analysis of the proposed approach based on the computation time. The computation time is evaluated by varying the number of clusters. The figures 7, 8 and 9 illustrate the performance in terms of computation time by varying the cluster values from 2 to 7.

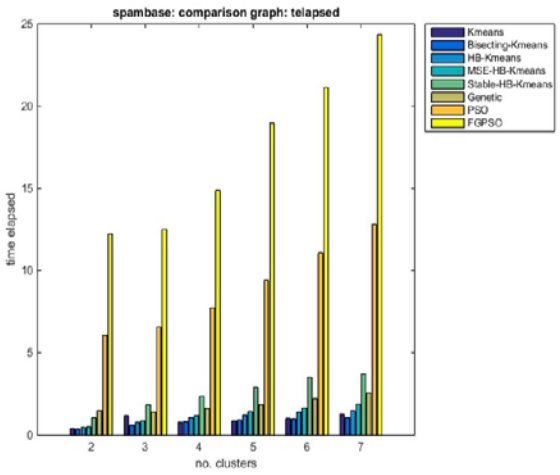


Fig.7. Comparative analysis in terms of computation time using dataset 1

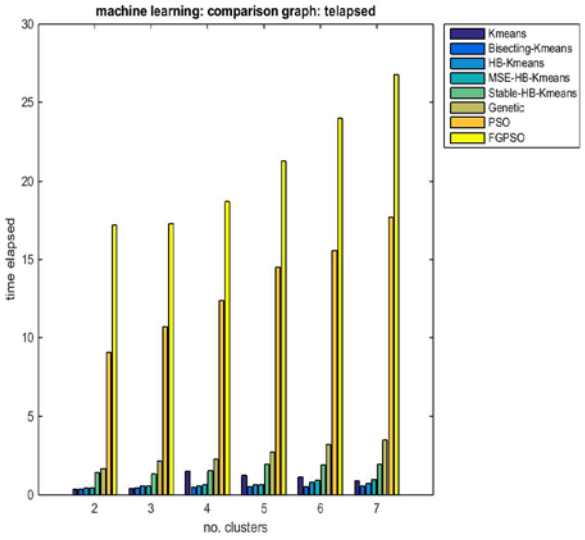


Fig.8. Comparative analyses in terms of computation time using dataset 2

The fig. 7, 8 and 9 represent the comparative analysis of the different algorithms for high dimensional data clustering using different datasets. This comparative analysis shows the analysis over computation time. According to the analysis from the graphs plotted above, it can be assessed that as the number of clusters increases, the computation time also increases for all the methods. In fig. 7, our proposed approach achieves the maximum computation time. When the number of cluster is 7, a computation time of 24 ms is achieved. In the same way in fig. 8 and fig. 9 also the maximum computation time is obtained for FGPSO when the total cluster value is 7. The time utilization is high for FGPSO because of the included constraints and

that expands the circle in the handling stage which results in increased time consumption.

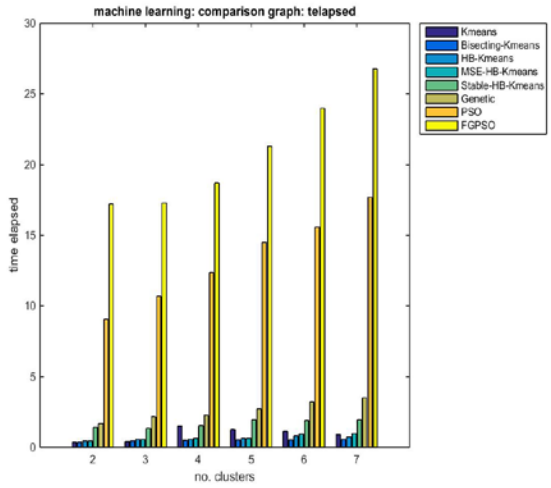


Fig.9. Comparative analysis in terms of computation time using dataset 3

## 5 Conclusion

The ultimate aim of this work is to develop a multi-objective optimization (MOO) technique by including different indices, like DB index, XB-index and sym-index. These three objectives have been used as fitness function for the proposed Fractional Genetic PSO algorithm (FGPSO) which is the hybrid optimization algorithm for clustering. The proposed multi objective FGPSO algorithm is implemented using MATLAB and the performance has been evaluated based on clustering accuracy and computational time. The proposed algorithm shows promising results.

*References:*

[1] Aparna K and Mydhili K Nair, "Comprehensive Study and Analysis of Partitional Data Clustering Techniques", *International Journal of Business Analytics*, Vol 2, Issue 1, pp. 23 – 38, January-March 2015.

[2] Aparna K and Mydhili K Nair, "HB-K Means: An Algorithm for High Dimensional Data Clustering Using Bisecting K-Means", Submitted for publication in *International Journal of Computational Science and Engineering*, Inderscience Publications.

[3] Aparna K and Mydhili K Nair, "CHB-K Means Algorithm: Incorporating Constraints to HB K-Means Algorithm", Submitted to *IETE Journal* of Research.

[4] I E Evangelou. DG Hadjimitsis, A A Lazakidou, C Clayton, "Data Mining and Knowledge Discovery in Complex Image Data using Artificial Neural Networks", Workshop on Complex Reasoning an Geographical Datal Cyprus, 2001.

[5] T Lillesand, Ralph W Keifer & Jonathan Chipman, "Remote Sensing and Image Interpretation", John Wiley & Sons. 1994.

[6] H C Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", John Wiley & Sons, 1972.

[7] M R Rao, "Cluster Analysis and Mathematical Programming", Journal of the American Statistical Association, Vol. 22, pp 622-626, 1971.

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

[9] JR Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Inc. Publishers, 1993.

[10] G Potgieter, "Mining Continuous Classes using Evolutionary Computing", Department of Computer Science, University of Pretoria, Pretoria, South Africa. 2002.

[11] Dharmendra K Roy and Lokesh K Sharma, "Genetic k-means clustering algorithm for mixed numeric and categorical datasets", International Journal of Artificial Intelligence and Applications, Vol.1, No.2, 2010.

[12] B. Everitt, S. Landau, and M. Leese, "Cluster Analysis", London: Arnold, 2001.

[13] A. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[14] A. Baraldi and E. Alpaydin, "Constructive feed forward ART clustering networks—Part I and II", IEEE Trans. Neural Networks, Vol. 13, No. 3, pp. 645–677, 2002.

[15] V. Cherkassky and F. Mulier, "Learning From Data: Concepts, Theory, and Methods", New York: Wiley, 1998.

[16] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient K-Means Clustering Algorithm: Analysis and Implementation", IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, 2002.

[17] Ujjwal Maulik and Sanghamitra Bandyopadhyay, "Genetic algorithm-based clustering technique", Pattern Recognition, Vol. 33 ,pp.1455-1465, 2000.

[18] Sanghamitra Bandyopadhyay, Ujjwal Maulik and Malay Kumar Pakhira, “Clustering using simulated annealing with probabilistic redistribution”, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 15, no. 2, pp. 269-285, 2001.

[19] Weihui Dai, Shouji Liu and Shuyi Liang, “An Improved Ant Colony Optimization Cluster Algorithm Based on Swarm Intelligence” Journal of software, Vol. 4, No. 4, 2009.

[20] Jayshree Ghorpade - Aher and Vishakha A. Metre, “Clustering Multidimensional Data with PSO based Algorithm”, Soft Computing and Artificial Intelligence, 2014.

[21] H.S Behera, Rosly Boy Lingdoh And Diptendra Kodamasingh, "An Improved Hybridized K-Means Clustering Algorithm (IHKMCA) For High dimensional Dataset & its Performance Analysis", International Journal On Computer Science And Engineering (IJCSE), Vol. 3, No. 3, pp. 1183-1190, 2011.

[22] Tulin Inkaya, Sinan Kayalıgil and Nur Evin Ozdemirel, “Ant Colony Optimization based Clustering Methodology”, Applied Soft Computing, vol. 28, pp. 301-311, 2015.

[23] Liyong Zhang, Witold Pedrycz, Wei Lu, Xiaodong Liu and Li Zhang “An interval weighed fuzzy C-Means clustering by genetically guided alternating optimization” Expert Systems with Applications, vol. 41, no. 13, pp.5960-5971, 2014.

[24] Spambase Data Set from http://archive.ics.uci.edu/ml/datasets/Spambase

[25] Localization data for person activity dataset https://archive.ics.uci.edu/ml/datasets/Localization+data+for+person+activity.

[26] Pen-Based Recognition of Handwritten Digits Data Set from “http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits”.