Asian Journal of Information Technology 15 (13): 2217-2227, 2016
ISSN: 1682-3915

A New Hybrid Algorithm for Finding Automatic Clustering in Unlabeled Datasets

[1]Komarasamy Ganeshan and [2]Amitabh Wahi
[1]Department of Computer Science and Engineering,
[2]Department of Information Technology, Bannari Amman Institute of Technology,
Sathyamangalam, India

Abstract: In data mining the clustering techniques is used for grouping a set of physical or abstract objects into similar objects. In this process, k-means algorithm is a major role to group the similar objects. The major issue of this algorithm is the user gives the number of clusters in priori as k value where as the final clustering results is ineffective. To avoid such a problem a new Multi Objective (MO) method Bat Modified Clustering Multi-Objective Optimization (BATMClustMOO) is proposed. This algorithm is a combination of Archived Multi-Objective Simulated Annealing (AMOSA) and Bat Algorithm (BA) is suggested which can partition the data into a suitable number of clusters k and then find the best cluster centroid automatically. The AMOSA acts as the local search and BA acts as the global search to fix the number of clusters and cluster centroid. Each cluster is splitted into many small hyper spherical sub clusters and the centroid of all small sub-clusters is fixed into a string that comprises the entire clustering. In order to verify the performance of the proposed algorithm the different benchmark datasets are taken from UCI repository. The experimental results show the proposed method is better than the existing methods.

Key words: AMOSA, BatMClustMOO, benchmark, clustering, multi objective, pareto-optimal

INTRODUCTION

Jain and Dubes (1988) described Data mining is a technique which extracts the large amounts of data to discover interesting patterns. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is a popular unsupervised technique which divides the input datasets into k regions based on similar or dissimilar properties where the value of k (number of clusters) may or may not be known in prior to the clustering process.

The main objective of clustering methods is to find a partition matrix P(X) of the given dataset of X. The X consisting of n patterns and x_1, x_2, ... x_n such that partition matrix in the Eq. 1:

$$P(X) = \sum_{k=1}^{k} \sum_{j=1}^{j} p_{kj} = n \quad (1)$$

Finding the suitable number of clusters from a dataset is an important factor in clustering methods. The partition matrix P(X) of size k×n, represented as P $[p_{kj}]$ where k = 1 to K and j = 1 to n where is a part of x_j to cluster C_k. Finding the suitable number of clusters from a dataset is an important factor in clustering methods.

The obtained results need to verify before convergence and also to validate the acquired partitioning, many cluster validity indices have been proposed in the literature. The standard approach of discover the number of clusters is to apply a given clustering algorithm for a range of k values and to estimate an assured validity function of the resulting partitioning in each case.

Bandyopadhyay and Maulik (2001) have experimented non-parametric genetic clustering for comparison of validity indices. The variable string length genetic algorithm is used when the number of clusters is not fixed to prior, method is called novel nonparametric clustering approach. Chromosomes are encoded various number of clusters. The crossover operator is considered for string length and fitness is used to measure the cluster validity index. This method suggests to fixing the optimum number of clusters using validity indices. Instead of the Euclidean distance method the Point Symmetry (PS) distance method is used for assigning data points to the corresponding clusters. The Sym-index measure is used to validate the clustering results. It supports irrespective of the cluster shape, size and also reduces the time complexity with the help of KD (K-Dimensional) tree.

Corresonding Author: Komarasamy Ganeshan, Department of Computer Science and Engineering,
Bannari Amman Institute of Technology, Sathyamangalam, India

Maulik and Bandyopadhyay (2002) described a simulated annealing based methods to fix the effective clusters using the cluster validity indices. The use of the various validity indices and clustering methods is automatically identifies the fixed number of clusters which is demonstrated experimentally for both artificial and real-life datasets.

Kim and Ramakrishna (2005) presented an analysis of design rules implicitly used in defining cluster validity indices and reviewed different existing cluster validity indices. The method of using cluster validity indices to choose the optimal number of cluster depends on the better clustering algorithm. The algorithm efficiency depends on various factors including the initial values, algorithm parameters, optimization process and assumptions regarding the cluster selections. The cluster shapes are depends on certain assumptions of validity measures. But many cluster shapes exist in the same dataset, the algorithm results are unsuccessful to obtain the same results.

Bandyopadhyay and Saha (2007) authors stated a new symmetry based genetic clustering algorithm which automatically finds the number of clusters and fixed partitioning given dataset. In order to assigning points to various clusters by used PS based distance instead of Euclidean distance. It is used PS based cluster validity index to determine of the validity of the equivalent partitioning. This study also describes a single objective genetic clustering technique Variable string length Genetic Algorithm with Point Symmetry (VGAPS) to identify the number of clusters efficiently.

Handl and Knowles (2007) described a MO clustering process called MO Clustering with automatic K-determination (MOCK) is implemented which out performs many single-objective clustering methods. The main aim of this method is to handle clusters with hyper spherical shape or "connected" but well-separated structures from the given datasets. This method is less successive in datasets overlapping clusters which do not contain any hyper spherical shape of clusters. The number of data points are many string length becomes higher where as the convergence speed is very less.

Wang and Zhang (2007) described the fuzzy logic based cluster validity indices that are used to find the number of clusters. The cluster analysis main objective is to categorize the groups of related objects and helps to determine the allotment of patterns and attractive correlations in large datasets.

Bandyopadhyay *et al.* (2008) proposed a new method called Archived MO Simulated Annealing (AMOSA) is an efficient MO description of the SA algorithm. The MO optimization is used when dealing with the real world issues where there are many objectives that should be optimized concurrently. The Simulated Annealing (SA) is used into a search technique for solving complex optimization problems which is the basis of statistical mechanics principles.

Yang (2011) researcher described a BA for MO optimization method. The Multi-Objective Bat Algorithm (MOBA) is validate against a subset of investigation functions and then it is useful to solve MO design problems such as welded beam design. MO optimization issues are more complex than single objective optimization to find estimated optimality fronts. In addition, the algorithm has to be modified to accommodate MO properly.

Saha and Bandyopadhyay (2012) stated several connectivity based cluster validity indices are introduced. The cluster validity index concept basis of connectedness of the clusters is used. This index is able to detecting the suitable partitioning from the datasets which is having clusters of many shape, size or convexity as long as they are well splitted.

Saha and Bandyopadhyay (2013) described a new MO clustering method is Genetic Clustering MO Optimization (GenClustMOO) proposed which can automatically split the data into an fixed number of clusters. All the cluster is divided into many hyperspherical subclusters and the centers of all sub-clusters are fixed in a string to signify the entier clustering. For assigning points to various clusters, these local subclusters are considered individually. For the purpose of objective function estimate, these sub-clusters are combined to form the variable number of global clusters. Three objective functions are used to validate the final number of clusters.

Arockiam *et al.* (2012) described a clustering methods and algorithms in data mining. In this study, to study the data clustering algorithm needs in all the applications by using k-means partitional clustering approach.

There are many validity indices are studied in the literature review to determine the validity of clusters is represented in the sequence of clusters in terms of the integrity. In order to use effective method for getting number of clusters is a challenging task in clustering process.

Multi objective optimization for clustering: The clustering process is a complex task because there is no specific dividing of the data exists for many datasets. Many clustering methods depends on only one criterion which reflects a single access of goodness of a partitioning the given datasets. The single cluster performance measure is seldom equally applicable for

various kinds of datasets with many characteristics. Hence, it is necessary to parallel process for many cluster performance measures that can take the various data characteristics. In order to solve these issues of clustering a MO optimization method GenClustMOO has been applied Saha and Bandyopadhyay (2012).

The MOCK is used to handle clusters with hyper spherical shape or "connected" but well-separated structures from the given datasets. This method is less successive in datasets overlapping clusters which do not contain any hyper spherical shape of clusters described Bandyopadhyay and Saha (2007)

The initial assignment of k value depends on the final clustering performance, to avoid such a problem proposed a new MO clustering technique called BATMClustMOO. This algorithm is a combination of AMOSA and BA, which can automatically partition the data into suitable number of k clusters and best cluster centroid.

Instead of data points of the cluster a new method BATMClustMOO is used for encoding of cluster centers based on the bat flow. It is used to identify the fixed number of clusters and the suitable partitioning from datasets with various types of cluster formations. A SA hybrid approach called BA Yang (2011) is used to identify the cluster centroid of each cluster based on the bat flow. Another method MO optimization approach AMOSA is used as the underlying optimization process. The multiple centers of each cluster are used to fix the number of clusters.

The entire cluster is divided into many non-overlapping small hyper spherical sub clusters and the centers of these sub clusters are fixed in a string to identify a particular cluster. Three search capability of AMOSA used for cluster validity indices are optimized concurrently Maulik and Bandyopadhyay (2002). First cluster validity indices used for total compactness of a particular partitioning, second used for total symmetry shows in a particular partitioning and third measure to get the degree of "connectedness" of a particular partitioning Bandyopadhyay and Maulik (2001).

The MO optimization methods generate a large number of non-dominated results on its last pareto optimal front. All of these results gives a way of dividing the particular dataset. The major issue is that sometimes the user need a single result from the non-dominated results is very difficult. This paper implements a new semi supervised scheme to get the single best result from the set of final pareto optimal solutions.

SA based AMOSA algorithm: The AMOSA method is a MO version of SA, many process have been included. It retrieves the process of an archive where the non-dominated results are stored. The restrictions are kept as the size of the archive which has two limits. One is a hard or strict limit mentioned by HL and second is a larger or soft limit mentioned by SL, where SL is greater than HL. The non-dominated results are stored in the archive and when they are produced.

In the non-dominated process, if few members of archive get dominated by the new results, then members are rejected. If some time, the size of the archive cross a specified value then the clustering process described as following.

In AMOSA, the initial point temperature is set to T_{max}. Then, first points (current-pt, or the initial result) are selected randomly from the archive. The current-pt make a new result called new-pt then find the objective function.

Next check the domination result of the new-pt with respect to the current-pt and the results in the archive. The new measure is called amount of domination Δdom (a,b) between two results a and b is defined by the Eq. 2:

$$\Delta dom(a,b) = \prod_{i=1, f_i(a) \# f_i(b)}^{M} \frac{|f_i(a)| - f_i(b)}{R_i} \quad (2)$$

Where, $f_i(a)$ and $f_i(b)$ is ith objective values of the two results, R_i is the corresponding value of the objective function gets from the those in the population and M is the number of objective functions. The results of new-pt, current-pt and the points in the archive various cases are arise as described follows.

Case 1: new-pt may be dominated by current-pt or it is non-dominating with belongs to the current-pt but few points in the archive dominate the new-pt (Saha and Bandyopadhyay, 2013). Assume new-pt is dominated by sum of k points (contains current-pt and points in the archive). This case is established in Fig. 1.

The points D to H suggest the content of the archive at any time, other points demonstrate the various cases that may arise with respect to the archive points. Where F represents the current-pt and B represents the new-pt. Then a quantity Δdom_{avg} is find using the Eq. 3.

$$\Delta dom_{avg} = \frac{\sum_{i=1}^{k} \left(\Delta dom_{i,new-pt}\right) + \left(\Delta dom_{current-pt,new-pt}\right)}{1+e} \quad (3)$$

The probability ratio (p_{qs} of new-pt and current-pt as mention in the Eq. 4:

$$p_{qs} = \frac{1}{1+e\left(\frac{\Delta dom_{avg}}{T}\right)} \quad (4)$$

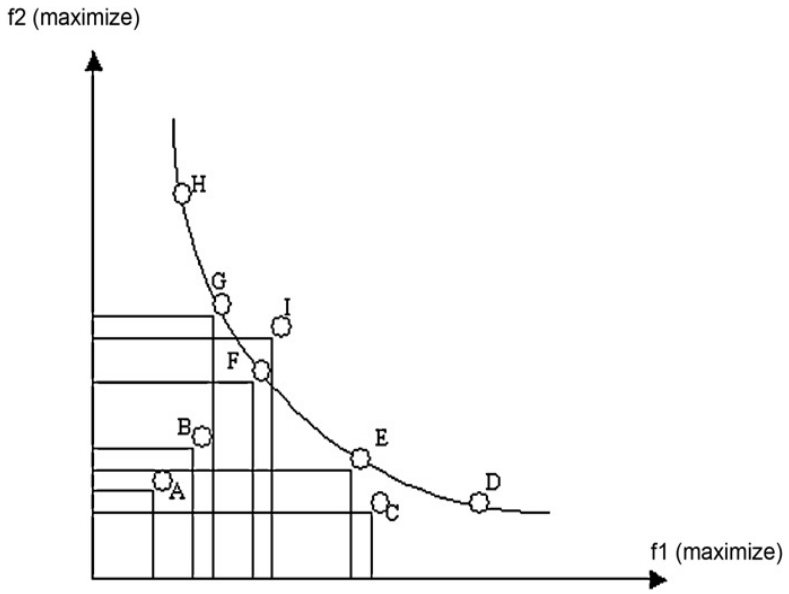

Fig. 1: Pareto-optimal front and unlike domination points

The $\triangle dom_{avg}$ represent the average amount of domination of the new-pt by (k+1 points, the current-pt and k points of the archive. It the value of k increases then the value of $\triangle dom_{avg}$ also increase. The dominating points are further missing from the new-pt are added to the value.

Case 2: Another case the current-pt nor points which is in the archive dominate the new-pt. Figure 1, the various points are accepted in the basis on the new-pt as the current-pt. The cases F represents the current-pt and E represents the new-pt, G represents the current-pt and I represents the new-pt, F represents the current-pt and I represents the new-pt. If any points in the archive which are dominated by new-pt, remove from the archive and next combine new-pt in the archive by Eq. 5. If archive size exceeds the SL, relate single linkage clustering to decrease its size to HL:

$$\Delta dom_{avg} = \frac{(\sum_{i=1}^{k}(\Delta dom_{i,new-pt}))}{k} \quad (5)$$

Case 3: The new-pt dominates the current-pt but k points belongs to archive dominate the new-pt . Fig. 1 shows the point A which represents the curent-pt and B represents the new-pt. Here the least variations of domination amount between the new-pt and k points, denoted by $\triangle dom_{min}$ where the archive is calculated. The current-pt is selected based on archive that corresponds to the minimum variations of probability by the Eq. 6. Another case, new-pt is selected as the current-pt.

$$probability = \frac{1}{1+\exp(\Delta dom_{min})} \quad (6)$$

Where, $\triangle dom_{min}$ is minimum of the difference domination amounts between the new-pt and k points.

The archive point is accepted only based on the residing of the non-linear shape of clusters. Any temperature temp all the cases are continued iter times. Temperature is reduced to $\alpha \times$temp, using the cooling rate of α till the minimum temperature T_{min} is reached and archive contains the final non-dominated solutions.

MATERIALS AND METHODS

Bat algorithm: Bats are one of the best nocturnal amphibians to fly in the night time. They are the only mammals which has wings and special capability of echolocation. Most micro bats are insectivores. The bats signalize the type of sonar called echolocation to identify the prey. In order to avoid obstacles and locate their roosting crevices in the dark location. Usually bats emit a loud sound pulse frequency and pay attention for the echo that bounces back from the immediate objects.

The BA steps to idealize the echolocation characteristics of microbats to use following approximate rules applied (Yang, 2011): all bats utilize echolocation to get the distance and they also know the difference between food or prey and background barriers move in some supernatural way.

Bats fly randomly with velocity v_i at position x_i with a constant frequency f_{min}, updated wavelength λ and loudness A_0 to search for prey. They often change the wavelength or frequency of their pulse emission and adjust the rate of pulse emission rϵ [0,1] depending on the proximity of their target. The loudness raises the ways from maximum A_0 to minimum constant valueA_{min}.

In order to create a new solutions by changing the pulse frequency, velocities and locations based on the Eq. 7-9:

$$f_i = f_{min} + (f_{max} - f_{min})\beta \quad (7)$$

$$v_i^t = v_i^{t+1} + (x_i^t - x^*)f_i \quad (8)$$

$$x_i^t = x_i^{t+1} + v_i^t \quad (9)$$

Bat algorithm:

Based on these approximations the Bat algorithm is described as follows:

Objective function f(x), x = (x_1, ... x_d)
Initialize bat population x_i = 1,2 ... n and v_i
Set the pulse frequency f_i at x_i
Initialize the rates r_i and the loudness A_i
While (t<Maximum number of iterations)
Generate new solutions by updating the pulse frequency, velocities and locations by the Eq. 7-9
If (rand>r_i)
Select a solution among the best solutions
Generate a local solution among the selected best solution
End if
Generate a new solution by flying randomly
If (<A_i and f(x_i<f (x*)))
Accept the new solutions
Increase r_i and reduce $A_i$values
End if
Rank the bats and find the current best x*
End while
Post process results and visualization

In this process, the tracing method is not used to estimate the time delay and distance between the bats. The algorithm is very simple and easy to apply in MO optimization purpose because it cannot be extended in multidimensional cases.

BATMClustMOO Algorithm: In the BATMClustMOO algorithm used to determine the number of clusters and effective cluster centroid. After getting the number of clusters, identify the best solution from the BA. After the cluster optimization, the three objective functions are used. They are cluster validity index based on symmetry: Sym-index, Connectivity based cluster validity index: Con-index and Euclidean distance based cluster validity index: I-index.

In the proposed algorithm, assume the generation counter t = 1; Initialize the population NP bats randomly select and each bat corresponding to a potential solution to the given problem; pulse frequency f_i and the initial velocities v_i, set pulse rate r_i and A_i

Alogrithm 1:

While the termination criteria is not satisfied or t<MG (Maximum Generation)
do
Choose T_{max}, T_{min}HL, SL, rand, iter, α, temp = T_{max} pulse rates r_i and loudness A_i
Objective function f (x), x = $(x_1, ... x_d)^T$
Initialize Archive and bat population x_i = 1,2, ... n and velocity v_i
Set the pulse frequency f_i at x_i
Current-pt random (Archive)
/* Select random solution from archive*/
While (temp<T_{min})
for (i=0; i<iter; i++)
new-pt = perturb (current-pt)
verify the domination status of new-pt and current-pt.
/*various cases of points selection*/
if (current-pt dominates new-pt) /*case 1*/

$$\Delta dom_{avg} = \frac{\sum_{i=1}^{k}(\Delta dom_{i,new-pt}) + (\Delta dom_{current-pt})}{k+1}$$

/* k = total number of points in the Archive which dominate new-pt≥ 0*/
Set new-pt as current-pt with probability

$$p_{qs} = \frac{1}{1 + e(\frac{\Delta dom_{avg}}{T})}$$

If (current-pt and new-pt are non dominating to each other)/ *case 2*/
Check the domination status of new-pt and points in the Archive
If (new-pt is dominated by k, (k≥1)) points in the Archive)/*case 2(a)*/

$$\Delta dom_{avg} = \frac{(\sum_{i=1}^{k}(\Delta dom_{i,new-pt}))}{k}$$

Set new-pt as current-pt with the probability

$$p_{qs} = \frac{1}{1 + e\left(\frac{\Delta dom_{avg}}{T}\right)}$$

If (new-pt is non-dominating w.r.t all the points in the Archive)/* case2(b)*/
Set new-pt as current-pt and add new-pt to the Archive
If Archive-size>SL
Cluster to HL number of clusters
If (new-pt dominates k, (k≥1)) points of the Archive)/*case 2 (c) */
Set new-pt as current-pt and add it to Archive
Remove all the k dominated points from the Archive
If (new-pt dominates current-pt)/* case 3*/
Check the domination statues of new-pt and points in the Archive
If (new-pt is dominated by k, (k≥1)) points in the Archive)/* case 3 (a) */

$$probability = \frac{1}{1 + \exp(-\Delta dom_{min})}$$

ΔDom_{min} = minimum of the difference domination amounts between the new-pt and k points.
Set point of the Archive which corresponds to Δdom_{min} as current-pt with probability = prob
Else set new-pt as current-pt
If (new-pt is non-dominating with respect to the points in the Archive).
/* case 3 (b) */
Select the new-pt as the current-pt and add it to the Archive.

If current-pt is in the Archive, remove it from Archive.
Else if Archive-size>SL.
Cluster archive to HL number of clusters.
If (new-pt dominates k other points in the Archive)/*Case 3(c)*/
Set new-pt as current-pt and add it to the Archive.
Remove all the k dominated points from the Archive.
End for
Temp = α*temp
End While
If Archive-size>SL
Cluster Archive to HL number of clusters.
Validate the clusters by using the Eq. 7.7, 7.12 and 7.13
If (rand>r_i) /* Apply Bat algorithm to optimize best number of clusters */
Select a solution among the best solutions
Generate a local solution around the selected best solution
End if
Generate a new solution by flying randomly
If (rand <A_i and $f(x_i)$<f(x*))
Accept the new solutions
Increase the rates r_i and reduce the loudness A_i
End if
Rank the bats and find the current best x* number of clusters.
t = t + 1;
End while

In the proposed algorithm, first define the termination criteria then initialize the archive. Next step of the algorithm is selecting the solution randomly from the archives. Based on the iteration the new point is obtained by perturbation. After finding the new point the dominant status new point over the current point is calculated by Eq. 3. If the new point has more dominant than the current point then the new point is set as a current point by the Eq. 5. The current point is selected based on archive that corresponds to the minimum variations of probability by the Eq. 6.

From the dominant status, if obtain non-dominant performance with each other then the dominant status between the new point and points in the archive. In this case, suppose if the non-dominant status is between the new point and all points in the archive then that new point set as a current point and added that point to the archive. Then, in the iterative manner calculating this dominant status between the points. If the new point is dominates the points in the archive then that k dominated points are removed from the archive. Once get the HL (Higher Limit) number of clusters, optimize the effective clusters to apply BA. Based on the bat echolocation choose a solution among the best solutions, next to generate a local solution around the selected best solution. Check for the higher limit that reached the loudness and frequency of each bat with current best solution. If it is reached, select the best number of clusters otherwise increase the termination criteria.

Objective functions used: In order to get the effective clustering the objective functions are used with proposed algorithm (Bandyopadhyay *et al*., 2008). For the purpose three diverse cluster validity indices are used to get effective clustering result:

- Sym-index: symmetry based cluster validity index
- Con-index: connectivity based cluster verification index
- I-index: euclidean distance based cluster verification index

The following are described briefly how the three validity indices are used to find the effective number of clusters.

Symmetry based cluster validity index: Sym-index: New clustering methods for symmetry based distance assess to gets the goodness of clustering on different partitions of a data set (Handl and Knowles, 2007). The new cluster validity function Sym is defined in Eq. 10 and 13:

$$Sym(k) = \left(\frac{1}{k} \times \frac{1}{\varepsilon k} \times Dk\right) \quad (10)$$

Where;

$$\varepsilon k = \sum_{i=1}^{k} Ei \quad (11)$$

$$Dk = \max_{i,j=1}^{k} \left\| \bar{c}_i - \bar{c}_j \right\| \quad (12)$$

$$E_i = \sum_{j=1}^{n_i} d_{ps}^{*}(\bar{x}_j^{-i}, \bar{x}_i) \quad (13)$$

Where:

$\bar{x}_i, \bar{c}_i$ = Euclidean distance between the points

D_k = The highest Euclidean distance between two cluster centers among all sets of centers

The Sym-index is considered three factors, 1/k, 1/εk and D_k. The 1/k increase as k decreases, this k value needs to be decreased for Sym-index. The factor 1/εk measure the total cluster within the symmetry. If the ε_k value is less achieve effective symmetrical structures. If it balances three factors are complementary in nature, surely gets the proper grouping of number of clusters.

Connectivity based cluster verification index: Con-index: Connectivity based cluster validity Indices to maintain the effective clustering. The Con-index are used to get well-separated clusters among all the clusters (Saha and Bandyopadhyay, 2012).

Consider the clusters represented by C_k, for k = 1, 2, ... k, number of clusters as k. The medoids of kth cluster is $\bar{m}_k$ is the lowest average distance to all the

other points in the cluster. The point which has the lowest average distance to all the points in the kth cluster is represented in the Eq. 14:

$$\min i\text{index} = \text{argmin}_{i=1}^{nk} \frac{\sum_{j=1}^{nk} d_e(\bar{x}_i^k, \bar{x}_j^k)}{nk} \quad (14)$$

Where, n_k is total number of points in the kth cluster and $\bar{x}_i^k$ is the ith point of the kth cluster. The Con-indexis defined in the Eq. 15:

$$Con = \frac{\sum_{i=1}^{k} \sum_{j=1}^{nk} d_{short}(\bar{m}_i, \bar{x}_j^i)}{n \times \min_{i,j=1 \wedge i \# j}^{k} d_{short}(\bar{m}_i, \bar{x}_j^i)} \quad (15)$$

Where, $d_{short}(\bar{m}_i \bar{x}_j^i)$ is the shortest distance along the relative neighborhood graph between the two points $\bar{m}_i$ and $\bar{x}_j^i$ the jth point of the ith cluster. Con-index can be find using two components, the denominator represents the two medoids points among a total of k clusters as in least shortest distance. The Con-index numerator represents the sum of the connectedness of particular partitioning clusters.

If the clusters are grouped effectively minimum distance between the medoids and any point of particular cluster is small, numerator has very small value. Thus Con-index gets the minimum value when clusters are connected as well as separated effectively.

Euclidean distance based cluster verification index: I-index: The connectivity based cluster validity Indices is to maintain the effective clustering. The Con-index is used to get well-separated clusters among all the clusters. It maintains the effective clustering. Cluster performance validation based on I-index denoted in Eq. 16-18 stated by (Bandyopadhyay et al., 2008):

$$I(k) = \left(\frac{1}{k} \times \frac{\varepsilon_1}{\varepsilon_k} \times D_k\right)^p \quad (16)$$

Where;

$$\varepsilon_k \sum_{i=1}^{k} \sum_{j=1}^{nk} de(\bar{C}_k, \bar{x}_j^k) \quad (17)$$

$$D_k = \max_{i,j=1}^{k} de(\bar{C}_i, \bar{C}_j) \quad (18)$$

Where:

$\bar{C}_j$ = The center of the jth cluster

$\bar{x}_j^k$ = The jth point of the kth cluster

nk = Sum of points present in the kth cluster

The value of k for which I-index takes largest value is taken as the suitable number of clusters. The index I is a combination of three factors are used, 1/k, ε1/εk and Dk. If the number of clusters k is increased it reduces I-index. Another factor ratio between ε_1 as given dataset and if ε_k decreases the value of k is linearly increased. Therefore I-index increases as ε_k decreases. Third term of I-index is $D_k$is used to measure the highest value between two clusters that may increases the values of k. The upper bound reach the maximum limit with each datasets.

From the three diverse clusters validity indices are used to get best number of clusters. An objective function based clustering algorithm tries to minimize (or maximize) a function such that the clusters that are obtained when the minimum/maximum is reached are homogeneous. Finally to apply BA to optimize best number of clusters from the dataset Yang (2011).

RESULTS AND DISCUSION

The results of proposed algorithm is evaluated with k-Means algorithm. The benchmark datasets are taken from UCI repository are described in Table 1. For the simulation purpose the Matlab 2012a Software is used to implement the proposed algorithms (Table 2).

The proposed BATMClustMOO algorithm has been compared and analyzed with the existing k means and GenClustMOO algorithms. Table 3 described the investigation methods that are much efficient with a maximum performance than the existing methods.

Accuracy comparison: Accuracy is defined only the proportion of the true results. It is a combination of both true positives and true negatives in the given dataset. Accuracy can be calculated by the Eq. 19:

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}} \quad (19)$$

Table 1: Description of datasets

Datasets	No. of instances	No. of attributes	No. of clusters
Wine	178	13	3
Iris	150	4	3
Vehicle	946	18	4
Glass	214	10	7
Liver	345	7	3
Wisconsin	699	10	2

Table 2: Confusion matrix

Actual result	Detected result
True positive	False negative
False positive	True negative

Table 3: Comparision table of KM, GenClustMOO and BATMClustMOO algorithm

Algorithm	Data sets	k means	GenClust MOO	BATMClust MOO
Accuracy (%)	Wine	65.16	97.19	98.31
	Iris	87.33	100	100
	Vehicle	67.02	98.51	99.06
	Glass	69.15	98.13	98.55
	Liver	68.11	98.26	99.24
	Wisconsin	72.1	99.14	99.14
Precision	Wine	0.65	0.972	0.983
	Iris	0.87	1	1
	Vehicle	0.67	0.985	0.985
	Glass	0.69	0.981	0.990
	Liver	0.67	0.983	0.984
	Wisconsin	0.72	0.991	0.991
Recall	Wine	0.65	0.971	0.983
	Iris	0.87	1	1
	Vehicle	0.67	0.985	0.985
	Glass	0.69	0.981	0.991
	Liver	0.68	0.981	0.986
	Wisconsin	0.72	0.991	0.991
F-measure	Wine	0.65	0.971	0.983
	Iris	0.87	1	1
	Vehicle	0.67	0.985	0.985
	Glass	0.69	0.971	0.982
	Liver	0.67	1	0.99
	Wisconsin	0.72	0.985	0.985

Figure 2 shows the simulation and analytical results of accuracy rate. It is observed that proposed BATMClustMOO algorithm automatically partition the given dataset into a suitable number of clusters using BA and AMOSA algorithms. For all the datasets proposed method gives better accuracy than existing methods like k means and GenClustMOO. The accuracy is considered as y axis and the number of datasets are considered as x axis. It shows that BATMClustMOO algorithm produces 99.15% maximum improvement in terms of accuracy for all the datasets.

Precision comparison: Precision is described as the fraction of a cluster that consists of objects of specified class. Another way of defining precision is probability (choosing at random) of objects from the specified class: precision is ratio of the true positives among the cluster by addition of true positives and false positive. This can be calculated by the Eq. 20:

$$\text{Precision}(i,j) = \frac{m_{ij}}{m_i} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (20)$$

Where, m_{ij} is sum of points are associated cluster i and j and m_i is sum of points are associated with cluster I. Figure 3 shows the precision and datasets comparison chart. It is observed that proposed BATMClustMOO algorithm gives effective number of cluster using multiple objective functions with BA. But existing k means and GenClustMOO algorithms gives less recall rate. The precision rate is considered as y axis and the number of datasets are considered as x axis. Hence, the proposed BATMClustMOO algorithm gives better precision rate than the exiting methods.

Recall comparison: Recall is defined as probability of related objects which is selected from the specified class. In other words recall is described as a combination of all objects that are grouped in to a specific class. Recall is a function of the proper classification of data (true positives), and misclassified data (false negatives). Assume cluster i with respect to class j, then recall can be find by Eq. 21.

$$\text{Recall}(i,j) = \frac{m_{ij}}{m_j} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (21)$$

Where, m_j is sum of points are associated with cluster j. Figure 4 shows the recall and datasets comparison. It is observed that proposed BATMClustMOO algorithm gives better recall rate with effective number of cluster using multiple objective functions. But existing k means and GenClustMOO gives less recall rate. Hence the proposed BATMClustMOO algorithm gives much better recall rate than the exiting methods.

F-measure comparison: F-measure is a combination of precision and recall that measures the extent to a cluster that contains only objects of a particular class and all objects of that class. The F-measure of cluster i with respect to class j is defined by Eq. 22 and 23.

$$\text{F-measure}(i,j) = \frac{2\times\text{precision}(i,j)\times\text{recall}(i,j)}{\text{precision}(i,j)+\text{recall}(i,j)} \quad (22)$$

Then;

$$F(i,j) = \frac{2PR}{P+R} \Rightarrow F_c = \frac{\sum_i |i| \times F(i)}{\sum_i |i|} \quad (23)$$

Where, every class i is associated a cluster j which has the highest F-measure, F_c represents the overall F-measure that is the weighted average of the F-measure for each class i and $|i|$ is the size of the classes.

Figure 5 shows the F-measure and datasets comparison. It is observed that proposed BATMClustMOO algorithm gives better F-measure rate with effective number of cluster using multiple objective functions. But existing KM and GenClustMOO gives less F-measure rate. Hence the proposed BATMClustMOO algorithm gives better F-measure rate than the exiting methods.

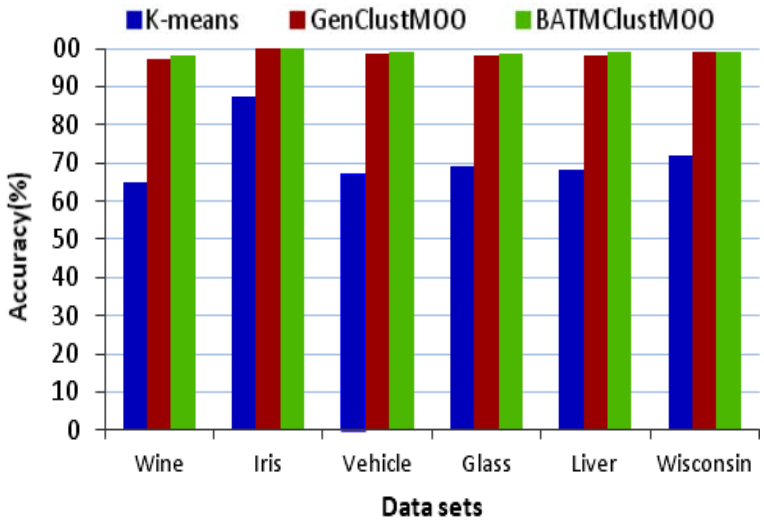

Fig .2: Accuracy comparison chart of BATMClustMOO algorithm

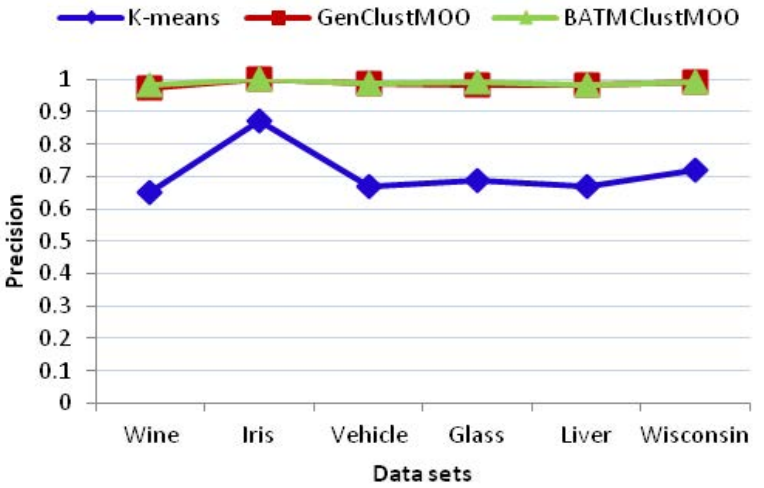

Fig. 3: Precision comparison chart of BATMClustMOO algorithm

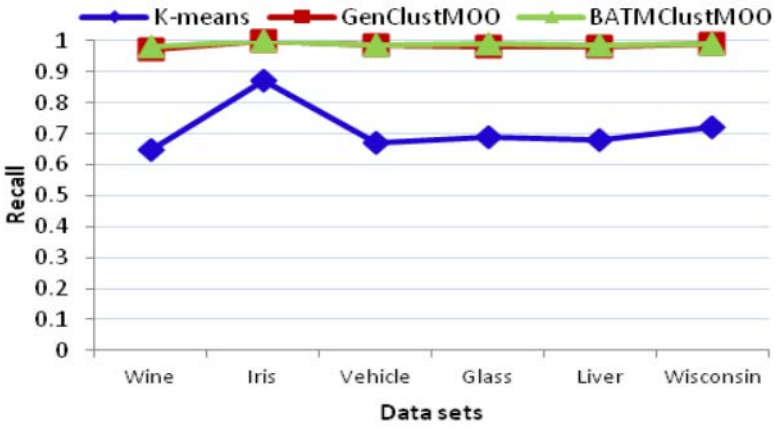

Fig. 4: Recall comparison chart of BATMClustMOO algorithm

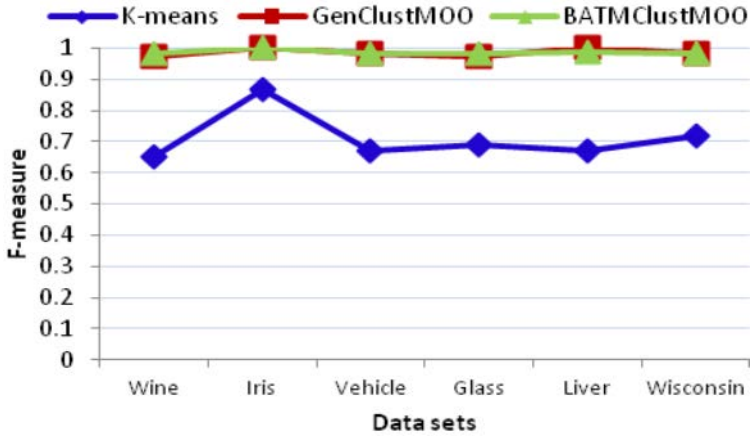

Fig.5: F-measure comparison chart of BATMClustMOO algorithm

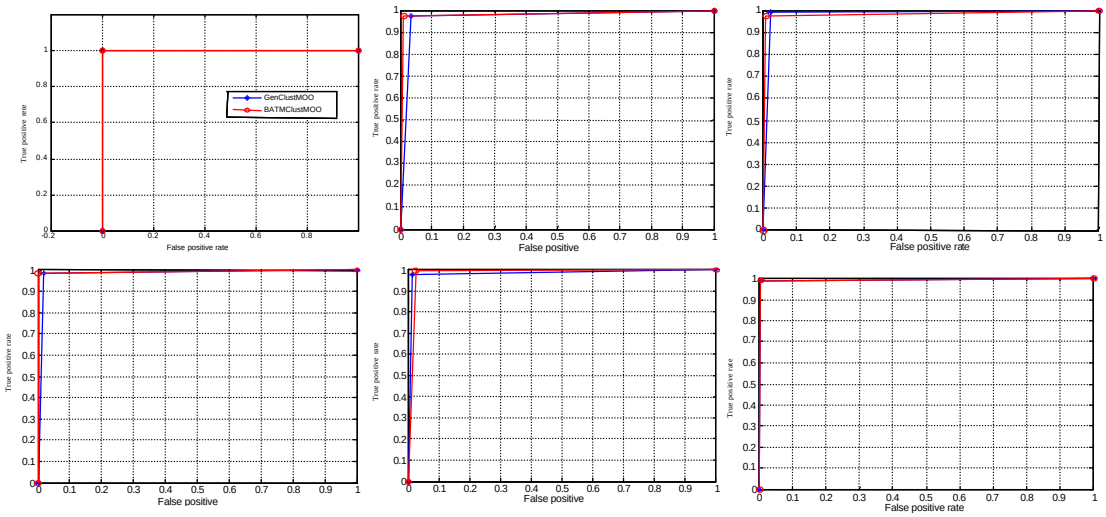

Fig. 6: a) ROC wine dataset; b) ROC iris dataset; c) ROC vehicle dataset; d) ROC glass dataset; e) ROC liver dataset; f) ROC Wisconsin dataset

Receiver Operating Characteristic (ROC) curve: The ROC curve is used in graphical representation between True Positive Rate (TPR) and False Positive Rate (FPR) of a given dataset. In the ROC curve x axis represent the FPR, it shows the clusters which have negative incorrect group or total negative group of given dataset. The y axis represents TPR, it shows the total positive group of given dataset.

Fawcett (2004) illustrated a ROC graphs, it is useful for classifying and visualizing their performance of an algorithm. It is commonly used in medical decision making, machine learning and data mining research community.

The ROC curve represents the performance of algorithm with respect to FPR and TPR respectively for all the datasets. The ROC curve resides (0,0) means each object of the cluster declared in negative class. If resides (1,1) then each object of the cluster declared in positive class and it resides (1,0) means clusters are ideal position.

The ROC curve having following conditions are applied: The ROC curve resides 1.0, it is perfect prediction, 0.9 states excellent prediction, 0.8 states good prediction, 0.7 states mediocre prediction, 0.6 states poor prediction, 0.5 states random prediction and <0.5 states something wrong in the clustering process. The above performance metrics are applied in all the proposed algorithm. In the proposed algorithm there are some linear improvements based on the cluster formation.

Figure 6 shows the ROC graphs for all the six datasets. It is very useful technique for represent the performance with FPR and TPR of all the datasets.

Figure 6a presents the Wine dataset chart, both the methods give higher TPR. In Iris dataset chart

Fig. 6b shows the both the methods gets similar result where as the proposed method exhibits little higher TPR. Similarly vehicle, glass and liver dataset projects the higher TPR like result of Iris dataset as represented in Figures 6c-e datasets, respectively. At last Wisconsin datasets both the methods gets merged and exhibits the higher TPR result are observed in Fig. 6f.

CONCLUSION

In this study, we purposed a new MO clustering technique BATMClustMOO algorithm is used for identifying the fixed number of clusters in a given dataset automatically. All the clusters are splitted into number of hyper spherical subclusters and various clusters are assigning into points, from this local points sub-clusters are separated. The objective function uses these sub-clusters which are properly merged into variable number of clusters. Three objective functions are used reflecting the total compactness which is divided based on the Euclidean distance, next reflecting sum of the symmetry clusters and reflecting connectedness of the cluster are considered. After obtaining number of clusters, three diverse cluster validity indices are used and also BA algorithm is used to fix the exact number of clusters from AMOSA for getting effective results.

In order to verify the performance of the proposed algorithm BATMClustMOO is compared with the existing method GenClustMOO. Finally to verify the quality of the algorithm using accuracy, recall, precision and F-measure parameters.

This study also needs to focus on how to reduce the time complexity without compromising cluster quality and parallel distributed implementation of clustering algorithms to reduce the time taken by the computational process. More experiments will be conducted with complex natural datasets with different features.

REFERENCES

Bandyopadhyay, S. and S. Saha, 2007. GAPS: A clustering method using a new point symmetry-based distance measure. Pattern Recognit., 40: 3430-3451.

Bandyopadhyay, S. and U. Maulik, 2001. Nonparametric genetic clustering: Comparison of validity indices. IEEE. Trans. Syst. Man Cybern., 31: 120-125.

Bandyopadhyay, S., S. Saha, U. Maulik and K. Deb, 2008. A simulated annealing-based multiobjective optimization algorithm: AMOSA. IEEE Trans. Evolutionary Comput., 12: 269-283.

Fawcett, T., 2004. ROC graphs: Notes and practical considerations for researchers. Mach. Learn., 31: 1-38.

Handl, J. and J. Knowles, 2007. An evolutionary approach to MO clustering. IEEE. Trans. Evol. Comput., 11: 56-76.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. Prentice Hall Inc., Englewood Cliffs, USA., ISBN: 0-13-022278-X, Pages: 320.

Kim, M. and R.S. Ramakrishna, 2005. New indices for cluster validity assessment. Pattern Recognit. Lett., 26: 2353-2363.

Arockiam, L., S.S. Baskar, L. Jeyasimman 2012. Clustering Methods and Algorithms in Data Mining: Review Asian J. Inform. Technol., 11: 40-44.

Maulik, U. and S. Bandyopadhyay, 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE. Trans. Pattern Anal. Mach. Intell., 24: 1650-1654.

Saha, S. and S. Bandyopadhyay, 2012. Some connectivity based cluster validity indices. Appl. Soft Comput., 12: 1555-1565.

Saha, S. and S. Bandyopadhyay, 2013. A generalized automatic clustering algorithm in a multiobjective framework. Appl. Soft Comput., 13: 89-108.

Wang, W. and Y. Zhang, 2007. On fuzzy cluster validity indices. Fuzzy Sets Syst., 158: 2095-2117.

Yang, X.S., 2011. Bat algorithm for multi-objective optimisation. Int. J. Bio Inspired Comput., 3: 267-274.