CHAPTER 4

AN OPTIMIZED K-MEANS CLUSTERING TECHNIQUE USING BAT ALGORITHM

This chapter introduces the new algorithm K-Means and Bat Algorithm (KMBA), for identifying the initial centroid of each cluster. The cluster analysis is one of the primary data analysis methods and KM algorithm is suitable for grouping a large datasets (MacQueen 1967). It is a partition based clustering which is to assign the number of clusters *k* and arbitrarily chooses the initial centroid of each clusters. In the PSO algorithm to minimize the inter cluster variance but the initial centroid fixed based on the user specified input described in the previous chapter, in order to avoid such an issues the KMBA is proposed. Many alternative of PSO approach and development exist in the literature, various new metaheuristic algorithms have been implemented (Cui & Cai 2009; Yang & Deb 2010; Yang & Deb 2012; Tang et al 2012; Yang 2013). Various approaches are implemented to discover the initial cluster center in KM algorithm (Khan & Ahmad 2004; Junling et al 2009; Ran Vijay Singh et al 2011; Mohammed & Wesam 2012).

The proposed algorithm finds the distance between each data object and centroid based on the echolocation behaviour of bat position and velocity. This method calculates the cluster center based on Bat (BA) algorithm (Yang 2010) and then it forms the cluster by using the KM algorithm. Using this approach the dependency of convergence on initial centroid is selected. The final clusters are formed by the minimal computational resources and

time. The experimental result illustrates the proposed algorithm gives the better result than the existing KM and BA algorithm.

4.1 INTRODUCTION

The KM algorithm is described in the first chapter, it has many issues. Such as lack of choosing the number of clusters, selection of initial centroid, inadequacy with non globular cluster, more time complexity, not support more number of objects and converge to sub optimal results. Selecting the number of clusters and initial centroids selection in advance is an essential in cluster analysis described by Fahim et al (2006). The effective cluster should exhibit two main properties like low inter-class comparison and high intra-class comparison. AhamedShafeeq & Hareesha (2012) stated KM algorithm with an intension of improving the cluster quality and to fix the optimal number of cluster.

Dong & Qi (2009) stated a new hybrid approach used for PSO and KM algorithm. It finds the cluster center for KM algorithm but the high speed of convergence which often implies a rapid loss of diversity during the optimization process, which inevitably leads to undesirable premature convergence. That is the drawback of PSO algorithm, to avoid such issues BA is applied in the proposed algorithm. The KMBA algorithm overcome such problem like finds the cluster centroid without the undesirable premature convergence and it does not get trapped into local optima.

4.2 SYSTEM MODEL

4.2.1 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 (Richardson 2008).

Their pulse properties are correlated with their hunting strategies, depending on the species. Bats are fly based by following the leading bat with the frequency modulated signals to search the path for further fly, while others more often use constant frequency signals for echolocation. Their signal bandwidth moderation is based on the number of bats in the group and often better by using more frequencies.

The BA steps to idealize the echolocation characteristics of microbats, to use following approximate rules applied (Yang 2010).

1. 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.

2. 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 $r \in [0,1]$ depending on the proximity of their target.

3. The loudness raises the ways from maximum A_0 to minimum constant value A_{min}.

In order to create a new solutions by changing the pulse frequency, velocities and locations based on the Equations (4.1) to (4.4).

$$f_i = f_{min} + (f_{max} - f_{min})\beta \tag{4.1}$$

$$v_i^t = v_i^{t+1} + (x_i^t - x^*)f_i \quad (4.2)$$

$$x_i^t = x_i^{t+1} + v_i^t \quad (4.3)$$

Based on these approximations the BA algorithm is described as follows.

1. Objective function $f(x), x = (x_1, \ldots . x_d)^T$
2. Initialize the bat population $x_i = 1,2, \ldots n$ and v_i
3. Set the pulse frequency f_i at x_i
4. Initialize the rates r_i and the loudness A_i
5. While (t < Maximum number of iterations)
6. Generate new solutions by updating the pulse frequency, velocities and locations by the Equations (4.1) to (4.3)
7. If $(rand > r_i)$
8. Select a solution among the best solutions
9. Generate a local solution among the selected best solution
10. End if
11. Generate a new solution by flying randomly
12. If $(rand < A_i \& f(x_i) < f(x^*))$
13. Accept the new solutions
14. Increase r_i and reduce $A_i$values
15. End if
16. Rank the bats and find the current best x^*
17. End while

18. 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 any application for computational purpose because it cannot be extended in multidimensional cases. In addition to this consideration the general frequency f in a range $[f_{min}, f_{max}]$corresponds to a range of wavelengths $[\lambda_{min}, \lambda_{max}]$. For example, a frequency range of 30 kHz, 500 kHz corresponds to a range of wavelengths from 0.9mm to 20mm.

4.2.1.1 Velocity and position vectors of virtual bats

The virtual bats are used in the simulation reason. The Bat positions x_i and velocities v_i in a d-dimensional search space are changed in the basis of the initial condition. The Equation (4.1) to (4.3) are given in the changed solutions x_i^t and velocities v_i^t at time step t, where it is derived from the uniform sharing to random vector drawn in $\beta \in [0, 1]$. The x^* is the best current global solution which is located after comparing all the solutions among n bats. As the product of $\lambda_i f_i$ is the velocity increment, use either λ_i orf_i to alter the velocity modify while fixing the other factor λ_i or f_i depending on the type of interest. In the process to use f_{min} = 0 and f_{max} = 100, depending on the domain size of the problem of interest.

Initially, each bat is randomly assigned to a frequency which is drawn uniformly from f_{min} and f_{max}. For the local search, once a solution is chosen among the current best solutions, a new solution for each bat is identified by using random walk based on the Equation (4.4)

$$x_{new} = x_{old} + \in A^t \tag{4.4}$$

where, $\varepsilon \in [-1,1]$is a random number

$A^t = A_i^t$ is the average loudness of all the bats at this time.

The moderation of velocities and positions of bats uses some alike measures to the procedure in the standard PSO as f_i basically controls the space and range of the association of the teeming particles. (Kennedy & Eberhart 1997). The BA can be measured as equal combination of the standard PSO and the concentrated local search controlled by the loudness and pulse rate.

4.2.1.2 Variations of loudness and pulse emission

Based on the iterations process the loudness A_i and rate r_i of pulse emission has to be changed. As the loudness usually decreases once a bat finds its prey, while the rate of pulse emission increases, the loudness can be chosen by any specific value. For example, A_0 = 100 and A_{min}= 1. If A_0= 1 and A_{min}= 0, simplify A_{min}= 0 then bat has just finds its prey and temporarily stop emitting the sound. Based on this action the Equation (4.5) and (4.6) functions to derive the result (Kirkpatrick et al 1983).

$$A_i^{t+1} = \alpha + A_i^t \tag{4.5}$$

$$r_i^{t+1} = r_i^0[1 - \exp(\gamma t)] \tag{4.6}$$

where, α and γ are constants

α is similar to the cooling factor schedule in the simulated annealing.

For any $0 < \alpha < 1$ and $\gamma > 0$ to use $A_i^t \to 0, r_i^t \to r_i^0$ as $t < \infty$.In order to reduce the number of iterations to use $\alpha = \gamma$ and $\alpha = \gamma$ = 0.9 in the simulation purpose. Initially, each bat should have various values of loudness and pulse emission rate and this can be produced by unsystematic. For

example, the initial loudness A_i^t and initial emission rate r_i^0 can be zero or any value $r_i^0 \in [0,1]$ using the Equation (4.5) and (4.6). The new solutions are changed in the basis of loudness and emission rates, which means that bats are moving towards the best possible solution. (Holland 1975)

4.2.2 Bat Clustering Algorithm

The cluster formation basic steps of the BA can be used with KM algorithm mentioned as follows.

1. Objective function set maximum iteration number = $iter\_max$
2. Initialize the bat population $x_i = 1,2, \dots n$ and v_i
3. Define pulse frequency f_i and x_i
4. Initialize the pulse rates x_i and the loudness A_i
5. Input cluster data object m
6. Set t=1
7. While ($t < iter\_max$)
8. {
9. for i = 1 to n do

   $N_r(x_i(t)) = \{j: \|x_j(t) - x_i(t)\| < r\}$

   where $N_r(x_i(t))$ is the data set containing in local space within r of x. $f(x_i) = |N_r(x_i(t))|/m$
10. for each bat i do
11. {

    $N_i(x_i(t)) = \{j: \|x_j(t) - x_i(t)\| < r\}$ and $f(x_i) < f(x_j)$

    for each bat $j \in N_i(t)$

Select bat j has highest $f(x)$ by the Equation (4.1) to (4.3)

Generate a new solution by flying randomly

if($rand < A_i$ & $f(x_i) < f(x_j)$

Accept the new solutions

Increase r_i and reduce A_i

12. }
13. t =t+1
14. }

The first step of this algorithm is to define the number of bats involving by objective function $f(x)$. These bats will be initialized as x_i and v_i. Define the pulse frequency f_i at x_i also initialize the pulse rate r_i and loudness A_i of the bats. Provide the input of cluster objects m. The iterations value (t) will be started from 1 for the clustering process of the BA algorithm, with this the maximum iteration value (iter_max) will be assigned. When the iteration value is less than the maximum iteration value of the clustering process will be done to produce the solution (by the way of for loop execution) using BA until the best solution is obtained. This algorithm is less efficiency, so to overcome such issues the KMBA algorithm is exploited.

4.2.3 KMBA Algorithm Steps

In the proposed KMBA algorithm, each bat initialize the cluster location particle with two basic tasks.

- The first task is to search the best number of clusters by using a discrete PSO approach based on Cooperative Particle Swarm Optimizer (CPSO) which is proposed by Van den Bergh & Engelbrecht (2004).

- The second task is to find the best set of cluster centers by using the KM algorithm according to the assigned number of clusters.

The pseudo code for KMBA Algorithm

1. Initialize the objective function $f(x), x = (x_1, \dots x_d)^T$
2. Assign the bat population x_i where $i = 1,2, \dots n$ and v_i
3. Set maximum number of iterations is iter_max
4. Consider t=1
5. Define pulse frequency f_i at x_i
6. Initialize bat position rand, pulse rates r_i and loudness A_i
7. While (t <iter_max)
8. Generate a new best solutions by modifiable frequency, for i=1 to n do
9. Updating velocities and locations or solutions
    a. initialize K center locations
    b. assign each x_i to its nearest cluster center A_i
    c. update each cluster center $x_i$and mean of all x_i that have been assigned as closest to cluster.
    d. calculate the distance value (d_{ij})

$$d_{ij} = \sqrt{\sum_{k=1}^{x}(x_{ik} - x_{jk})^2} \quad (4.7)$$

    e. if the value of Euclidean distance has converged (near together)

    then return center locations

else go to step b

10. if (rand >r_i)
11. select a solution among the best solutions
12. generate a local solution around the selected best solution
13. End if
14. Generate a new solution based random fly
15. if($rand < A_i \& f(x_i) < f(x^*)$)
16. accept the new solutions
17. increase the rates and reduce the loudness
18. End if
19. rank the bats and find the current best x^*
20. Increment t
21. End while
22. Post process results and visualization
23. Get the solution from the above bat algorithm and fix the initial centroid;
24. Repeat
25. Assign each data item $d_i$to the cluster which has the closest centroid;
26. Calculate the new mean of each cluster;
27. Until convergence criterion is met.

This algorithm specifies the number of bats through objective function as $f(x), x = (x_1, \ldots x_d)^T$ and initialize the bats as x_i= (i = 1,2…n) and v_i,then set the maximum number iterations to find the initial clusters and consider the value as one. After that define the pulse frequency f_i then initialize the pulse rate r_i and loudness A_i.While the iteration value is less than the maximum iteration the function will get start to update the velocities, locations and sounds of all the bats. The *k* initial clusters centre locations will be initialized after assigning each bats of x_i to the nearest initial cluster centre $c_i$locations. Each cluster center updated by c_i as the mean of all x_i that have been assigned as closest to it. Calculate the Euclidean distance d_{ij} by using the Equation (4.6). If the value of Euclidean distance has converged then return center locations else do the step of assigning each bats to the initial cluster centre once again.

The random value (rand) of the other bats (whose value of Euclidean distance has not been converged) is greater than the pulse rate. Selecting the solution among the best solutions generates a local solution around the selected best solution. The function ends after generating a new solution by random fly of the bats. If random value is less than the pulse rate $f(x_i) < f(x^*)$then the new solution will be accepted. The pulse rate will be increased than the loudness rate. At last find the best bat among the $x_i$number of bats. Then increase the initial set value and continue the process until it gets the specified initial cluster centres. After finding the initial cluster centroid group the dataset by the KM algorithm. The KMBA algorithm flow chart is illustrated in Figure 4.1.

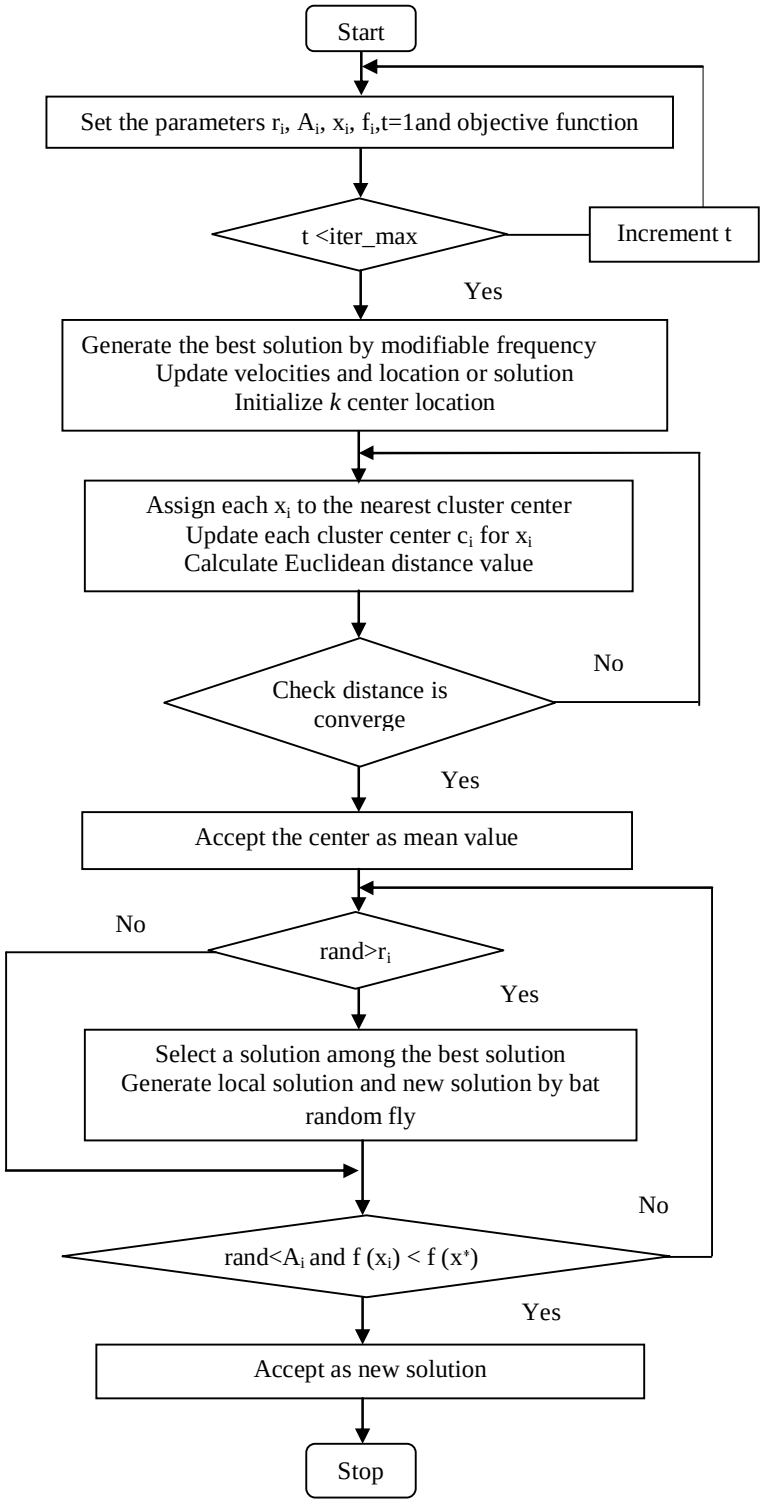

Figure 4.1 Flow chart for KMBA algorithm

4.3 RESULTS AND DISCUSSION

This section elucidates the experiments that are conducted to evaluate the robustness of the KMBA algorithm for initial centroid selection of the cluster. The comparison results of proposed KMBA with KM and BA algorithms are displayed in Table 4.1.

Table 4.1 Comparision table of KM, BA and KMBA algorithms

Algorithm	Data sets	KM	BA	KMBA
Accuracy (%)	Wine	65.16	92.13	**93.25**
	Iris	87.33	93.33	**96**
	Vehicle	67.02	92.12	**97.02**
	Glass	69.15	93.25	**95.32**
	Liver	68.11	91.01	**92.17**
	Wisconsin	72.10	93.13	**94.13**
Precision	Wine	0.65	0.92	**0.93**
	Iris	0.87	0.93	**0.96**
	Vehicle	0.67	0.92	**0.97**
	Glass	0.69	0.93	**0.95**
	Liver	0.67	0.90	**0.92**
	Wisconsin	0.72	0.93	**0.94**
Recall	Wine	0.65	0.92	**0.93**
	Iris	0.87	0.93	**0.96**
	Vehicle	0.67	0.92	**0.97**
	Glass	0.69	0.93	**0.95**
	Liver	0.68	0.91	**0.92**
	Wisconsin	0.72	0.93	**0.94**
F-measure	Wine	0.65	0.92	**0.93**
	Iris	0.87	0.93	**0.96**
	Vehicle	0.67	0.92	**0.97**
	Glass	0.69	0.93	**0.95**
	Liver	0.67	0.90	**0.92**
	Wisconsin	0.72	0.93	**0.94**

4.3.1 Performance measures

The proposed algorithm is evaluated using performance measures such as accuracy, precision, recall and F-measure with six benchmark datasets are described in chapter 1.

Accuracy comparison

Accuracy is defined only the proportion of the true results, it can be calculated by the Equation (1.6) which is mentioned in chapter 1.

Figure 4.2 illustrates the accuracy versus datasets comparison results. Based on the initial centroid of each cluster the KMBA gives better accuracy for all the datasets, but KM and BA gives less accuracy. The proposed method smartly increased from 85% to 91% of accuracy for all the datasets.

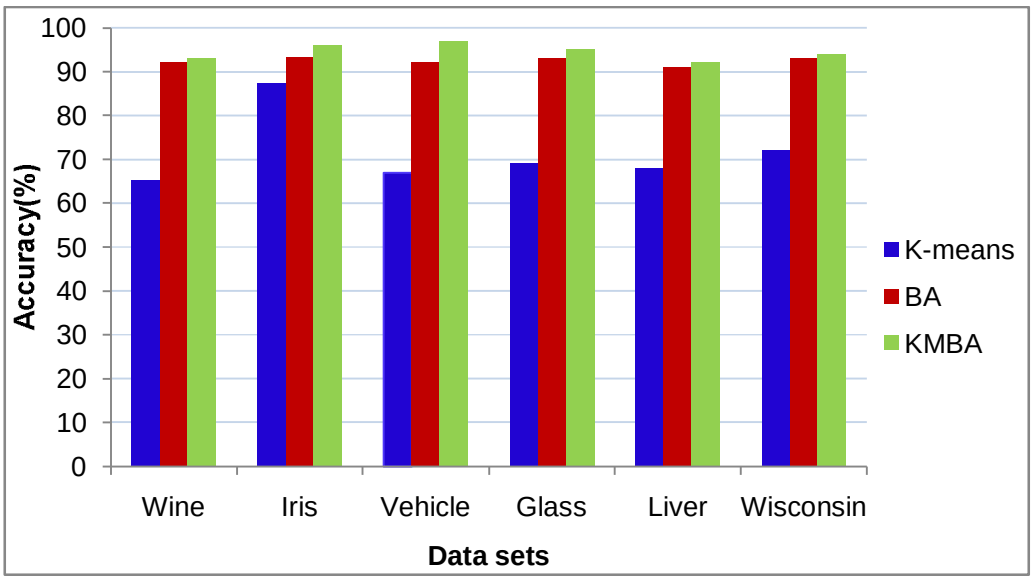

Figure 4.2 Accuracy comparison chart of KMBA algorithm

Precision comparison

Precision is estimating the ratio of the true positives among the cluster, it can be calculate by the Equation (1.7) which is discussed in chapter 1. Figure 4.3 displays the comparison among the datasets and precision. The KM algorithms give less precision rate in all the datasets, because it takes more iteration to compare with KMBA methods. The BA method has less precision rate. The KMBA produces more precision rate based on the initial centroid selection, it converge its minimum number of iterations.

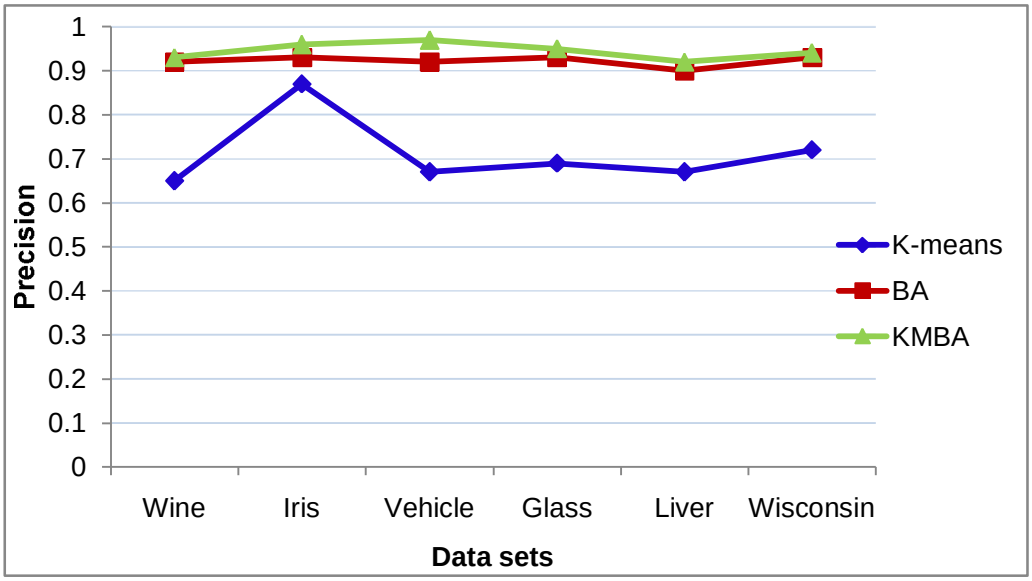

Figure 4.3 Precision comparison chart of KMBA algorithm

Recall comparison

Recall is desfined as a combination of all objects that are grouped in to a specific class, it finds using Equation (1.7) which is discussed in chapter 1.

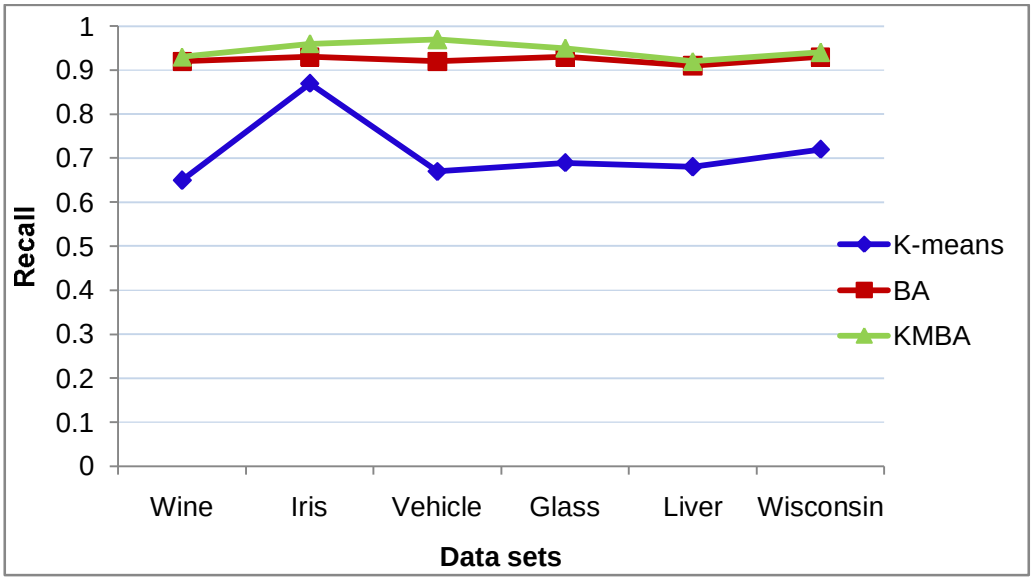

Figure 4.4 Recall comparison chart of KMBA algorithm

Figure 4.4 illustrates the recall versus datasets comparison results. Based on the initial centroid of each cluster the KMBA gives better recall rate for all the datasets, but KM and BA algorithm gives less recall rate.

F-measure Comparison

F-measure is a combination of precision and recall that measures the cluster that contains only objects of a particular class it finds by the Equation (1.9) and (1.10) which is discussed in chapter 1.

Figure 4.5 display the F-measure chart for all the six datasets, it clearly shows KMBA gives better performance than other methods. The KMBA algorithm F-measure is linearly increased based on the precision and recall measures. Form the experimental results it can be observed that KMBA algorithm gives better result than KM and BA algorithm.

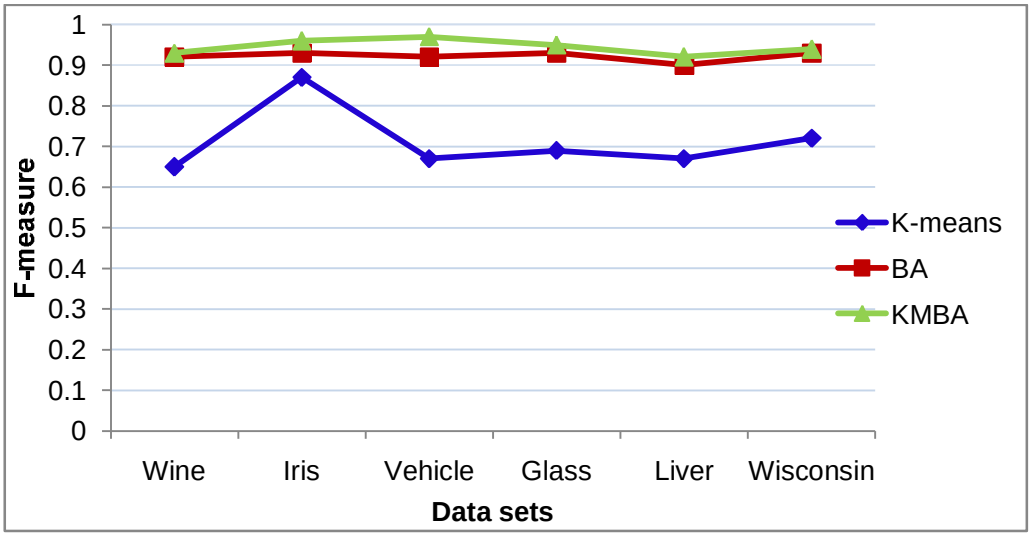

Figure 4.5 F-measure comparison chart of KMBA algorithm

ROC Curve

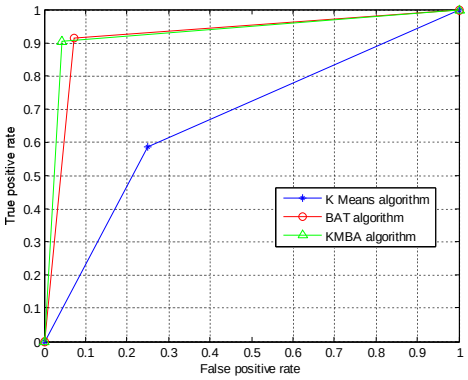

Figure 4.6 a Wine dataset

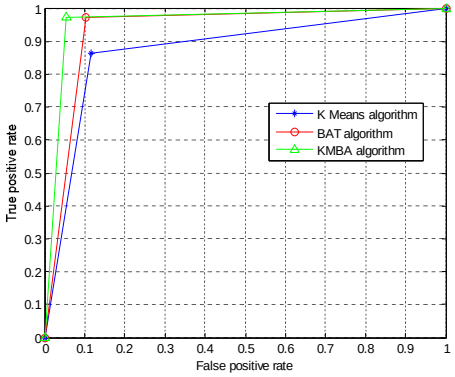

Figure 4.6 b Iris dataset

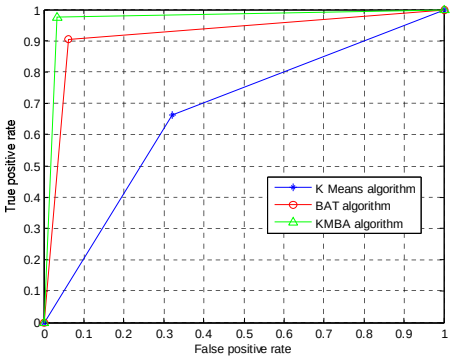

Figure 4.6 c Vehicle dataset

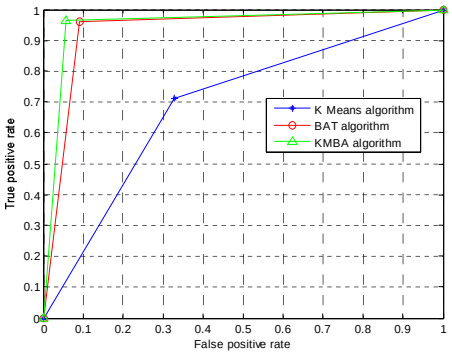

Figure 4.6 d Glass dataset

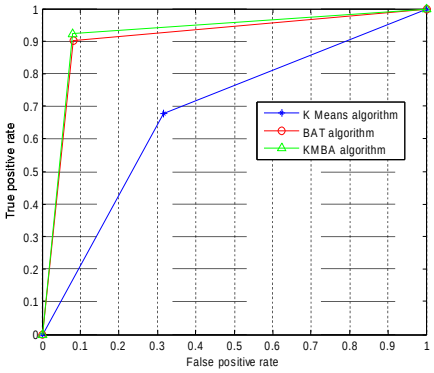

Figure 4.6 e Liver dataset

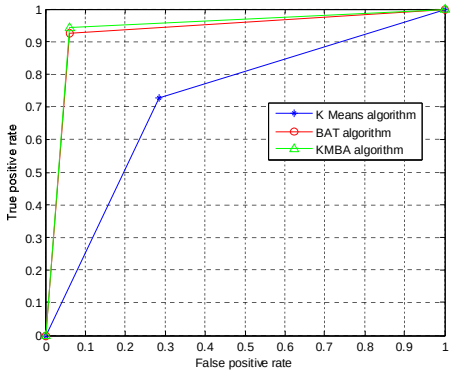

Figure 4.6 f Wisconsin dataset

Figure 4.6 ROC comparison graph of KMBA algorithm

Figure 4.6 shows the ROC graphs for all the six datasets. It is very useful technique to represent the performance with FPR and TPR of all the datasets, it can select certain conditions which are described in chapter 1. In the ROC curve x axis represent the FPR, shows the clusters which has the negative incorrect group or total negative group of given dataset. The y axis represent TPR, it shows the positive incorrect group or total positive group of a given dataset.

Form the above experimental results it can be observed that proposed KMBA algorithm gives better results. Hence KMBA gives effective centroid for all the clusters in the basis of bat flow automatically.

4.4 SUMMARY

In this chapter to address the problem of finding most informative features in clustering is initial centroid of each cluster using a new hybrid algorithm KMBA. It finds the distance between each data object and centroid based on the echolocation behaviour of bat position and velocity easily. In order to evaluate the performance of proposed algorithm the meta heuristic

algorithms are used in the robustness of the proposed approach. Six benchmark datasets are used to complete this task, in which KMBA gives better result than KM and BA algorithm. The proposed algorithm has outperformed in all the datasets. Hence the proposed algorithm cluster centroids may not be the optimal results, because in every run it selects random initial clusters. It can be converged into local optimal solutions. Therefore a new hybrid algorithm is required to find the global optimal solutions.