Nature-inspired Clustering Methodology
4 Nature-inspired Clustering Methodology

The involvement of computation process in every possible activity of people has made our society more data intensive. The information of an individual has become a necessary element or an important part of the human life, and hence the requirement to analyze those information for further use is an essential matter. The human analysis of its own stored data is either done by classifying or by clustering the group of data. Now classification is one of the most vital activities of the human being, and this is an essential technique to recognize any object, and once it recognizes that object then it can learn that particular subject or item. So, the classification technique is directly related to the human learning process. The process of clustering is also related to the human learning process where most similar items are grouped together, so that the process of prediction or learning or recognition of one item in a group is almost similar for other items. In practical sense the clustering algorithms are used to access the numerical information of each item or pattern and then organize them into a group. So, patterns within a cluster are more similar to each other than that of the patterns belonging to different clusters. In this chapter an automatic clustering method is proposed which is based on the bat optimization algorithm. The proposed method is compared with other standard clustering algorithm and the experimental analysis is performed over benchmark and synthetic dataset.

4.1 Introduction:

The act of clustering involves the partitioning of similar pattern into an improved homogeneous group which is known as cluster. One of the most essential activities of human being is clustering. The human clustering process is typically influenced by the society in which he/she is grown up, and it also differs on the basis of their educational status. In some cases different individuals may identify a single pattern as a similar one to a cluster, or identify that pattern as dissimilar one to that cluster. The features to recognize or detect a pattern is generally from one to many and based on those features human or machine categories a pattern to a particular cluster. Now it is very natural if the features list of a pattern is long, then it may take time for human individual to put that particular pattern into a class, whereas the algorithm can do it much faster and in a most efficient way. In case of long feature list with different
dimension of data, the human individual must experience enough to create a cluster from those data, but one single clustering algorithm can do it in most consistently and reliably. The data analysis individual must look for a pattern matrix in the data to put that pattern in such cluster, where almost all patterns are same, and then that individual must evaluate its own clustering procedure to judge the clustering quality. Now if the clustering algorithm performs the job of clustering then it is easy and also time saving for data analysis individual to analyze the performance of that clustering as well as quality of cluster created by that clustering algorithm. So the necessity of developing more improved clustering algorithm is always in the requirement list in various research fields.

The process of clustering is important for exploratory analysis to reveal the internal relationship among the data. The clustering algorithms mainly developed through four stages are:

a) Feature selection: In this phase of data clustering the data are presented to the algorithm by the type, number, scale and label. Some of these information on data set are really helpful to the clustering process. The feature selection is the process of finding the correct subset of data and its relative description to be processed through the clustering. Now the feature extraction generally involves the transformation of selected features of data so that it can be used by the clustering algorithm for proper partitioning.

b) Pattern proximity: In order to analyze the proximity, various types of techniques are available. The most commonly used one is the Euclidean distance measure which is actually helpful to find the similarity between two patterns, but this kind of approach is generally threshold based procedure. Various distance measuring techniques are available to find the similarity between two selected patterns and this various distance measure procedure may result to different shapes of the cluster.

c) Cluster formation: This phase is the result of previous one. The proximity measurement on the patterns is responsible to create different group of data set where each of those group consists of most similar patterns within it. Most of the clustering algorithms related to some kind of similarity measurement techniques. The selection of the proximity measurement technique is an important matter for any clustering algorithm and once it is chosen then it helps that algorithm in partitioning process which needs to be mathematically well defined problem.

d) Clustering validation: The validity assessment of clustering method is the technique
to determine the goodness of procedure output. Different validity measurement procedures are available and this phase leads to a different cluster. The final result of clustering may be affected by the parameter identification, presentation of input patterns or feature selection. So it is important to correctly measure the output of the clustering algorithm for determination of meaningful output. Individual cluster is valid, if the pattern within a cluster is not appeared in other cluster. Statistical hypothesis processes are applied to test the validity measurement and confidence level can be calculated from that. The validity measurement is of three types, they are, 1) External validity measure generally involves in comparison of the newly formed cluster and its previously formed cluster. 2) The internal validity measurement involves in each iteration phase of the clustering process to test the appropriateness of the internal cluster. 3) The relativeness test is performed between two clusters to test the relativity between them.

The clustering process is involved as a vital technique in various fields, ranging from engineering (e.g. machine learning, artificial intelligence, pattern recognition, mechanical engineering and electrical engineering), computer science,( e.g. image segmentation, web mining, spatial database analysis, textual document analysis, biomedical data processing etc.) , medical data analysis, earth science, microbiology, economics and various field of social sciences. In most of these above mentioned fields very less amount of information on the data are available, and hence the assumptions on data is done during the process of clustering. The strategy of clustering is also known as an unsupervised learning or learning without a teacher. Clustering technique is essential key factor for many fields, and hence the various development or modification is observed in clustering methods. It is observed that optimizing various standard problems based on nature-inspired algorithms attracts the researchers from various fields to implement that optimization method in the clustering algorithm for finding an improved result in clustering. Nature-inspired algorithms like particle swarm optimization (PSO), ant-colony optimization(ACO), cuckoo search(CS), firefly algorithm, BAT etc. are well known and clustering method based on these nature-inspired optimization algorithm can be found in the various literatures. Now in contrast to various types of development of clustering methodology, the main objective of this chapter is to provide a comprehensive and systematic experimental analysis based description of the influential and standard as well as some automatic nature inspired important clustering algorithm. The remainder of this chapter is organized as follows.
In the first phase some standard clustering method as well as automatic nature-inspired algorithm based clustering techniques are described, and those techniques are also used for experimental purpose. In the second phase most recently proposed BAT clustering method and its relative optimization procedure are explained. Then in third phase the comparative analysis and experiment result, discussion over various standard benchmark data set is proposed. Finally chapter is concluded in fourth phase.

4.2 Clustering Methods:

Data clustering algorithm is broadly divided into three categories: overlapping, hierarchical and partitional clustering methods. In the process of clustering, the patterns in the same cluster are similar to each other and patterns are not similar for different clusters. Now the validation or the judgment of the similarity and dissimilarity is needed to be measured in a clear and meaningful way. So for clustering, given data set must consist of \( X = \{ x_1, x_2, x_3, \ldots, x_n \} \) number of patterns and each of this pattern consists of \( x_m \) number of attributes or features where, \( x_m = (x_{m1}, x_{m2}, x_{m3}, \ldots, x_{mi})^T \in \mathbb{R}^i \) is said to be a feature of \( m^{th} \) pattern and \( i=(1,2,3,\ldots) \) is the dimension. A pattern can be differentiated with others based on its own attributes or features, and a pattern can be represented by data matrix \( x_{mi} \) with \( i \) dimensional row. Now for partitional and hierarchical clustering process there are some common considerations and they are:

> In case of partitional clustering the algorithms tries to find a \( C = \{ C_1, C_2, C_3, \ldots, C_k \} \) partition of \( X \) data set, where \( k=\{1,2,3,\ldots\} \) is the number of classes. The clustering is done in such a way that the similarity of patterns in \( C_k \) is maximum, and patterns of other cluster is not similar with the compared one. In case of partitional clustering the algorithm must maintain the following properties:

1) Each cluster must contain one pattern, so, \( C_i \neq \emptyset \) where, \( \forall i = \{1,2,3,\ldots,K\} \)

2) Two clusters always maintain dissimilarity between them. \( C_i \cap C_j = \emptyset, \forall i \in \{1,2,3,\ldots,k\} \)

3) Each pattern must be assigned a class.

The data set are partitioned into \( k \) number classes and there should be one fitness function to evaluate the performance of the algorithm.

> The hierarchical clustering is of agglomerative in type for which it either performs the clustering in the bottom-up approach or in top-down approach. The top-down or divisive approach of clustering starts with every element and then forms the
large clusters in every iteration. Now this approach constructs a tree like format of every partition \( X, H=\{H_1, H_2, \ldots, H_N\} \) where \((Q<N)\), such that

\[
C_i \in H_i \text{ and } C_j \in H_j \text{ where } C_i \cap C_j = \emptyset \text{ and } i, j = \{1, 2, 3, \ldots\}
\]

In hierarchical clustering algorithm, the main advantage is that the algorithms do not require any prior declaration of cluster number, and it is free from initial conditions. The one of the most disadvantage is that the elements or pattern assigned to one cluster is not movable to other one, so the assignment of pattern to a cluster is static in type. The hierarchical algorithms perform clustering on the basis of its proximity matrix. The results of the algorithm are represented through the dendrogram. The whole data set is represented through the root of the dendrogram, so in the bottom each of the leaf node represents the data object. Now in the dendrogram the intermediate node represents the similarity based on which the groupings can be performed. The dendrogram height usually expresses the distance between the clusters. The dendrogram can be broken at different level to get the different clusters.

Hierarchical algorithms are mainly classified as agglomerative methods and divisive methods. The agglomerative clustering is initiated with \( C \) number of clusters, and each of them includes one data object. The general agglomerative clustering consists of following procedures:

1. **Begin:**
2. Initiate with \( N \) number of Cluster.
3. Calculate the proximity matrix.
4. Search the minimum distance:
   \[
   D(C_i, C_j) = \min \{ C_m, C_l \} \text{ where, } m < l, l<N \text{ and } m \neq l
   \]
   Now, \( D(C_i, C_j) \) is a distance function in the proximity matrix and \( D(C_i, C_j) \) are the two clusters which can be combined to generate a new cluster.
5. Proximity matrix is updated with the new distance that is calculated in step 4.
6. Goto step 2 and Step 3 until the convergence of the algorithm.

Based on the literatures available over different agglomerative algorithm the distance calculation of the hierarchical algorithms is divided into two types Single linkage and complete linkage. Single linkage is based on the method of nearest neighbour and the distance is calculated between the two closest pattern of different clusters. The
Complete linkage is based on the furthest distance between two patterns of different clusters. Other than single and complete linkage algorithm, there are several other development of hierarchical algorithm like average linkage, group linkage, median linkage etc.

On the other hand the partitional clustering technique is used to obtain at-least one single partition of collections of patterns in a data set. The iterative process is required to improve the performance of clustering, and it is also based on some fitness or cluster validity measurement. This chapter represents some of these algorithms and also compares their performance with the automatic BAT clustering algorithm. The algorithms are:

a. DBSACN;  b. K-Means; c. Exception Maximization

The DBSCAN algorithm works on the density of an item based on neighbourhood of that item. Density based methods consider that clusters are dense sets of data items separated by less dense regions. The clusters are derived from different type of density based function like multivariate Gaussian, t-distribution, Euclidean etc.

Now if the distribution of a given data set are known, then assignment of a pattern to a cluster actually involves the parameter modification. Now if posterior probability $M(C_i)$ is known previously for cluster $C_i$, $i=(1,2,3.....K)$, where $K$ is assumed to be known, as because this is not an automatic clustering scheme. So, The conditional probability is $M(p|C_iT_i)$, where , $p$ is the pattern and belong to the data set and $T_i$ is the parameter and the value of this parameter is unknown. So the mixture density probability is

$$M(p|T) = \sum_{i=1}^{k} M(p|C_iT_i)M(C_i),$$

where $T = (T_1, T_2,.......N)$ and

$$\sum_{i=1}^{k} P(C_i) = 1.$$

Bayes theorem for calculation of probability is performed to assign a pattern to a cluster. The Exception Maximization (EM) algorithm is based on the maximum likelihood estimation. The EM algorithm divides the data set patterns $P_j$ into two parts, one part consists of the attributes of the pattern $p_j^n$ and another part consists of information about missing pattern $p_j^m$ and $p_i^m = (p_{i1}^m, p_{i2}^m,..... p_{ik}^m)$, where, the value of $p_{jm}$ is in between 0 and 1. So it is possible to decide whether $p_j$ belongs to the missing attribute category or not, and hence the maximum likelihood is calculated as follows:
Now, $M(C_i) = \text{the probability of assignment of an item or pattern to a cluster.}$

$t_i = \text{the parameter (} t^1, t^2,.... t^H) \text{ and modification of this is performed in each iteration, it is also treated as a convergence parameter.}$

The EM algorithm suffers from the possibility of convergence into a local optimum. The algorithm is also non-automatic clustering process and it requires the initialization parameters settings.

The K-means algorithm algorithm relies on the squared error method which is targeted to reduce the cost function. In general the k-means algorithm didn't initiate the proper number of clusters trying to improve the cluster by moving the patterns from one cluster to another, moving the patterns in such a way that the cost or fit function is reduced. The distance between the patterns and the cluster center is considered as a cost function.

Let $P = \{p_1, p_2, p_3, ... , p_i\}$, where $i = (1,2,3.....n)$ be the set of $n$ d-dimensional points clustered into the $K=(1,2,3....,m)$ number of clusters, $C = \{c_1, c_2, c_3,....k\}$. K-means algorithm finds a cluster in such a way that the sum of all squared-error is minimum.

Let

$$ SSE = \sum_{k=1}^{m} \sum_{i=1}^{n} (p_i - \mu_k) $$

Where, $p_i \in C_k$, and the main objective of the K-means algorithm is to reduce the SSE and produce a unique set of clusters. The algorithm initially starts with $k$ number of clusters and then assigns the patterns in such a way that the sum of squared error is reduced.

4.3 Automatic BAT Clustering:

BAT algorithm is an optimization algorithm, and it is developed from the inspiration of echolocation characteristics which is observed in micro-bats for searching of the prey. The conventional bat algorithm is categorized into nature inspired optimization algorithm which already proves its efficiency in the respective field. In this chapter the concept of recently proposed bat optimization algorithm is employed to design the automatic bat algorithm based clustering method. A new rule based merging approach is adopted for merging the similar cluster which provides an optimal number of clusters.
Experimental results on several standard benchmark data sets indicate that the automatic bat clustering algorithm with rule based merging concept provides a possible way to partition linear and nonlinear benchmark data set. The proposed method is compared with other standard algorithm over a nine numbers of artificial and real life benchmark dataset. It is observed from the experiment that the algorithm is faster and convergence is improved than that of the compared one.

Nature inspired or evolutionary algorithms are population based optimization algorithms, and are developed from the inspiration of natural concept. Numerical relationship is derived from the natural concept and these numerical formulae are used to simulate various phases of concerned natural life. Nature-inspired optimization algorithm is a solution to those combinatorial problems which are intended to solve not only optimal or near optimal solution but to find a good solution. Meta-heuristic methods are developed for finding a solution of combinatorial optimization problem in a discrete search space. The nature-inspired or evolutionary algorithms belong to the meta-heuristic category, as because each iteration of the evolutionary algorithm provides a different solution from the previous iteration, and in each iteration tries to provide a better solution for every candidate. One of the most common approaches used in these population based meta-heuristic algorithms is the fitness dependent modification on parameters of the population yielding possible solution to a given problem. Hence the solution is moved towards the better solution in the search space. Now nature-inspired, meta-heuristic population based algorithms are applied for the clustering to achieve better performance. In most of the cases the swarm intelligence based algorithms are used to perform the clustering.

Clustering is the process of partitioning the unlabeled dataset into a meaningful group of similar type of data. The successful group consists of most similar item which is evaluated by various processes of measuring the similarity between two patterns. The successful grouping creates a meaningful cluster and the purity of the corresponding cluster is maximum. The usage of nature-inspired optimization algorithm in clustering the dataset became most popular. Recent years many researchers have turned their attention towards the algorithms inspired from natural life, in quest to solve tough optimization problems ever more quickly. There are methods like ant colony optimization (ACO), firefly algorithm, particle swarm optimization (PSO) etc. which have extensive use in the recent past. The fact that these kind of meta-heuristic
algorithms are not constrained to solve only a particular type of problem, but having the potential to be applied to a diverse array of problems, and this has made them more popular.

These algorithms have been tested on a popular optimization problem namely the Traveling Salesman Problem[1]. This problem requires that a salesman is able to traverse a number of cities before returning to the starting point in the minimum possible time and also with minimum distance traversed. One analogous problem to this is the vehicle routing problem [2]. This also requires that a set of vehicles are able to serve customers situated at different places before returning to the depot by using the minimum number of vehicles and traversing the minimum possible distance. The ACO technique finds extensive use in this problem as this algorithm is reportedly one of the fastest available optimization technique. As these kinds of problems involve clustering, the ACO can display categorization problems if the number of clusters happens to be large. Document clustering is one such problem where categorization errors are often experienced. In those cases the sole potential of the ACO is not enough, but other algorithms like the K-means algorithm have to be incorporated in addition to the ACO for solving the categorization errors [1]. The combination of ACO and the K-means algorithm has been seen to work properly on 20 Newsgroup data [1]. Data clustering is an important problem that involves the grouping of data having a large amount of similarity in terms of their attribute values, into several classes where one class of data is distinctly different from the members of another class [3]. Data clustering has found its application in several fields including biology, statistics and more importantly data mining [3]. The K-means clustering algorithm is the most suitable, and it is also capable of being run under unsupervised solution domains [4]. This is exciting aspect about nature-inspired algorithms. These can work efficiently under unsupervised environments. Artificial bee colony technique has also been used in the domain of data mining, and has been reported to be a considerable improvement over the other available techniques [5]. The main algorithm proposed here, the bat algorithm, has found considerable attention and is a potential improvement over the traditional techniques. The bat algorithm can also be used to solve optimization problem where solution has many folds or more precisely the multi-objective optimization problems such as multiprocessor scheduling [7]. As mentioned before, this algorithm is also applicable to a large number of optimization problems. In this internet era where on-line
banking, smart shopping and social networking are on a rampage, often the users fall victim of phishing websites and end up furnishing their sensitive information to the miscreants. Some factors have been determined that are expected to represent a phishing website, and the bat algorithm has been applied on such data sets to optimize the solution in detecting the fraudulent website. The bat algorithm has been seen to achieve a success rate of 70% in detecting the phishing websites [6].

Clustering technique has played a significant role in the widespread field of science and technology [8-13]. In order to meet the requirement of different applications [15-17], various clustering algorithm have been proposed in the literature and the new clustering algorithm continues to appear. In most of the cases the implementation of the nature inspired algorithm in clustering technique belongs to the hierarchical clustering categories, and hierarchy of clustering consists of partition of data set which is known as cluster. Clustering algorithms are categorized into the partitioning clustering and agglomerative hierarchical clustering [14, 18]. Hierarchical clustering techniques are organized data in the nested sequence of group. The hierarchical clustering algorithm is represented by the tree like format [18]. The partition clustering algorithms are approached from wide range of field, like, expectation-maximization algorithms [20], graph theory [21], artificial neural network [22,23,24], evolutionary computation [25,26,27], swarm intelligence[28] etc. In the partition algorithmic problem, a k-cluster is to be found from a group of n-patterns in a d-dimensional space. The k number of clusters consists of most similar patterns [19]. In some cases the algorithms of partitioning clustering mechanism may be initiated with an unknown number of clusters or an exact number of cluster. In the hierarchical clustering the main idea is to build a binary tree of data that successively merges similar group of data points. In this procedure the most popular one is the agglomerative hierarchical clustering algorithm [29,30]. Several modified version of the algorithm like PSO algorithm [31], kernel based method of clustering algorithm etc. is also developed in the recent years. The kernel based clustering algorithm measures the similarity between two patterns based on the defined kernel, and this type of algorithms does not use the conventional procedure of clustering which depends on the Euclidean distance matrix. The kernel maps any two data pattern into higher dimensional feature space where data set is linearly separable [32,33]. The most useful example of this type of method is the support vector machine (SVM) [34-37], kernel base K-means [38] and fuzzy C-means [39].
In the present work, the proposed method is devised to apply the bat algorithm in solving automatic data clustering (without the prior knowledge of cluster center) problems. The proposed method considers bat population as a cluster center. In the bat optimization algorithm the position of the bat changes on the basis of an equation and the same equation is used here to change the cluster centre. The proposed method of clustering is based on these cluster centres and in this method a rule based merging technique is incorporated to remove the null cluster as well as the same technique is used to remove the frequent similar type of cluster. The proposed method is compared with four-number of standard clustering algorithms and it is tested over nine benchmark data sets.

The contribution to the present work is summarized as follows:

1. The bat optimization algorithm is incorporated to develop an automatic clustering algorithm. The proposed method introduced a bat algorithm based automatic clustering scheme for grouping complex and linearly non-separable datasets, without the prior knowledge of the number of manually occurring groups in the dataset.

2. A new rule driven cluster merging technique is introduced to judge the similarity between two partitions. There are four rules to judge the similarity between two independently selected groups, and all of these rules are theoretically explained in the corresponding section of this chapter. The statistical measurement is done on every partition, and similarity identification on this measurement is based on the defined rules and one threshold value.

3. The SSE of Euclidean distance between cluster items and its corresponding cluster center is calculated to track the performance of each iteration. The performance of the clustering is reported on the basis of clustering accuracy measurement.

The performance comparison between the proposed method and the other four standard methods is performed in order to establish the efficacy of present method over compared one. The method is tested over several standard real life benchmark datasets. The effectiveness of the data dimensionality growth in the performance of the proposed algorithm is studied with the synthetic dataset.

4.3.1. BAT Optimization Algorithm:

The bat optimization algorithm [41,42] is actually inspired from the special characteristic of bat as the bat echo-location characteristics [40]. In order to maintain the
proper presentation of proposed method, the outline idea of bat behavior and bat optimization algorithm are required.

4.3.2. Bat Algorithm:

The concept of bat echolocation characteristic is inspired to develop the bat optimization algorithm. The procedure of natural hunting characteristic is utilized to design the optimization concept. Some general rules are considered for designing the optimization method and it is as follows:

1. The echolocation characteristics of bat is considered here as the measurement procedure of distance between two objects. It is also considered that the bat also knows the differences between the echoes of prey and obstacles.

2. In the algorithm it is considered that the bats are flying with the velocity $p_i$ from a position $s_i$ with a minimum frequencies $f_{\text{min}}$ and varying wavelength $\lambda$ and loudness $L_0$ to search for the prey. The bats are capable of adjusting the frequencies $f_i$ and also the rate of frequencies $r \in (0,1)$ and the adjustment depends on the successful or unsuccessful searching of prey or for the type (small or large) of prey.

3. The emitted sound of bat varies for performing different activity in their social structure. The large loudness is represented by the $L_0$ and minimum loudness is represented by the $L_{\text{min}}$.

The rules are formed in order to design the optimization algorithm from the concept of bat hunting procedure. The algorithm not considered the complex scenario where two or more bats target their sonar on same object. In any type of implementation, any range of wavelength is considered and it can be adjusted on adjusting the frequencies.

The conventional bat optimization algorithm didn’t specify the way to consider the time delay and three dimensional topography. To maintain simplicity, the algorithm considers range of frequencies corresponds to the range of wavelength. It is also considered that the frequencies must belong within the range of minimum and maximum frequencies. These consideration leads to the BAT optimization algorithm.

The bat $P_i$, moves from a position $S_i$ with a velocity $V_i$ and frequency $F_i$. The algorithm calculates the next bat position using following formulas:

$$ F_i = f_{\text{min}} + (f_{\text{max}} - f_{\text{min}}) \beta \quad (1) $$

$$ V_{(i+1)} = V_{(i-1)} + (P_i - P) * F_i \quad (2) $$

$$ P_{(i+1)} = P_{(i-1)} + v_{(i+1)} \quad (3) $$
where, $\beta \in (0,1)$ and $P_*$ is the best position of the bat and the best solution of the current iteration is generated by comparing the current one with other. In the bat optimization algorithm the minimum frequency ($f_{\text{min}}$) starts with zero and maximum ($f_{\text{max}}$) is 100. The frequency range is totally depended on the problem. Initially each bat is assigned with randomly generated position, frequency and velocity but each assignment is uniform within its range. If the algorithm is able to find the local best solution among the all best solution, a new solution for each bat is generated locally using random walk.

$$P_{\text{new}} = P_{\text{old}} + \gamma L^{t}$$ (4) Where, $\gamma \in (-1,1)$

and $L^{t}$ is the average loudness of all bats at a specific time. The main convergence parameters $L_{i}$ and $r_{i}$ are modified on each iteration. The loudness actually decreases as the distance between bat and prey reduces and the rate of pulse increases. The loudness is adjusted using following formula:

$$L_{i}^{(t+1)} = \alpha L_{i}$$ (5)

and the pulse rate is calculated as follows:

$$r_{i}^{(t+1)} = r_{i}^{0} \left[ 1 - e^{(-\lambda t)} \right]$$ (6)

where, $\lambda$ and $\alpha$ are the constants and it is always $0 < \lambda < 1$ and $0 < \alpha < 1$. The loudness and rate pulse is adjusted if the new solution is improved.

4.3.3. BAT Algorithm based Clustering:

The concept of bat optimisation algorithm is embedded in the proposed method to partition a linear or non-linear dataset. The proposed method is an automatic clustering process because the process is initiated without the prior information of actual number of clusters in the dataset. The proposed method is based on the movement of cluster centre which is moved to its best position in every iteration and thus the patterns are transferred in between the clusters to form a cluster with almost similar type of data patterns. The bat optimisation procedure is introduced to find an accurate cluster centre for similar type of patterns. The performance of the each iteration is evaluated on the basis of clustering accuracy measurement and the same is used as one of the stopping criteria for this method.

Echolocation based hunting procedure of bat is simulated in bat algorithm and this optimisation technique is incorporated in clustering process to find the accurate cluster centre. The patterns of a dataset are moved in between the cluster on the basis of this cluster centre. In the proposed method the clustering process is designed on the
AutoMatic Bat algorithm based Clustering

1. Begin:
2. Initiate the algorithm with a random generation of bat population, frequency and velocity.
3. Perform initial clustering using randomly generated parameters.
4. Calculate the sum of square error of Euclidean distance for each cluster.
5. While (i < MCN):
   6. Perform following calculation for each cluster:
      i) Select one cluster centre and its corresponding cluster.
      ii) If selected cluster is empty then pass to next cluster or else modify randomly generated bat population, frequencies and velocities according to the eqn. (7), (8) and (9).
      iii) calculate pulse rate of each population using eqn.(13)
   7. If the cluster is the best solution among all local solution then modify that bat population using eqn.(5).
8. Perform the simulation of clustering with newly generated value.
9. Now, Calculate the average sum of square error of Euclidean distance of clusters.
10. If the New average sum of square error of Euclidean distance is better than old one then accept all newly modified value of each bat population otherwise discard the new modification.
11. Merge two cluster if they consist of same elements.
12. Calculate the Classification error percentage and display the clustering result and i = i+1.
(2) In bat optimization algorithm the selection of a best bat position is simulated through the average calculation of all the patterns in a cluster. In every iteration the best cluster is evaluated on the basis of clustering accuracy.

(3) The rate of pulse and the loudness are interrelated in bat optimization algorithm. But, here the rate of pulse is fixed to a constant value. The minimum frequency is set to a random value and maximum one is set to the maximum number of clusters found in each iteration.

The bat positional movement which is considered in the clustering process as a movement of cluster centre is calculated as follows:

\[ Fc_i = fc_{min} + (fc_{max} - fc_{min}) \alpha \]  \hspace{1cm} (7)

\[ V_{i+1} = V_i + (C_i - C^*) Fc_i \]  \hspace{1cm} (8)

\[ C_{i+1} = C_i + V_{i+1} \]  \hspace{1cm} (9)

Where, \( fc_{min} \) and \( fc_{max} \) are the minimum and maximum frequency and \( \alpha \) is the constant value in Eqn.7. In Eqn. 8 the cluster center is \( C_i \) where best cluster center is \( C^* \) and the velocity is indicated as \( v_i \). If the rate of pulse is less than a randomly generated value then the bat position movement is calculated as follows:

\[ C_{i+1} = C_i + \gamma L' \]  \hspace{1cm} (10)

where, \( \gamma = \) random value \( \in \{0,1\} \) and the loudness is set to a fixed value 0.001.

4.3.3.1. Clustering quality measure:

The square error criterion to evaluate the clustering algorithm performance measure is the most commonly used strategy. The basic intention of this approach is to obtain a partition among all other clusters which minimizes the square error. The performance of automatic bat clustering algorithm is evaluated using sum of square error of Euclidean distance between the each elements of a cluster and its corresponding cluster head. Suppose that the given n patterns of d dimension is partitioned into k clusters \( \{Cl_1, Cl_2, \ldots, Cl_k\} \) and each cluster \( Cl_k \) is assigned to \( n_k \) pattern. Each of this cluster is the result of any one population \( P \). The sum of euclidean distance between each element and cluster center is calculated as:

\[ d_{ji} = \sum_{j=1}^{k} \sqrt{(Cl_{ji} - Ch_{ji})^2} \]  \hspace{1cm} (14)

Now the average error of Euclidean distance for \( n_p \) number of pattern in cluster is as
follows:

\[ E(c) = \sum_{j=1}^{k} \frac{d_{ij}}{n_p} \]  

(15)

Now, the eqn.15 creates a single row matrix and the final sum of squared error of Euclidean distance is as follows:

\[ SSE_{eucl} = \sum_{j=1}^{N} E(c)^T \cdot E(c) \]  

(16)

4.3.3.2. Rule based merging technique:

A rule based merging technique is embedded into the bat algorithm based clustering. The rule based merging is used to find the similarity and non-similarity between any two number of clusters. The merging of clusters are performed after the bat algorithm based clustering. In bat algorithm based clustering several empty clusters and small fragments of clusters are created at the end of clustering.

So, in order to get rid of these empty clusters and small fragments of clusters the merging technique is adopted after the clustering phase. In the proposed method the clustering phase consists of two sub phases i.e. prior data processing phase and bat algorithm based clustering phase. In data processing phase the levels of each and every pattern of a dataset is separated and these levels of every pattern is maintained in a separate list. Now, at the time of bat clustering phase the patterns are shifted from one cluster to another cluster and hence, the index of the pattern changes accordingly. After clustering the levels are arranged, and separately clustered according to the changed index. So, in the proposed method levels of each pattern creates the clusters separately.

Now, in rule based merging technique any two number of clusters, \( C_i, C_j \) are selected for merging on the basis of following properties:

1. \( C_i, C_j \neq \varnothing \), where \( \forall (i, j) = \{1,2,\ldots,k\} \)
2. \( C_i \cap C_j = \varnothing \)

Now, after merging of \( C_i \) and \( C_j \) cluster, a new cluster \( C_N \) is formed which consists of levels of the pattern. Now, the standard deviation is calculated as follows:
\[ \varphi_{CN} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - \mu_i)^2} \quad (17) \]

where, \[ \mu_i = \frac{\sum_{i=1}^{N} X_i}{N} \quad (18) \]

\( X_i \) = label of a pattern \( P_i \), \( N \) = total number of label in \( C_N \).

In a similar way the standard deviation \( C_i \) and \( C_j \) cluster is \( \varphi_{ci} \) and \( \varphi_{cj} \) also calculated in a similar way. In the proposed method the following rules are applied for taking decision on similarity between \( C_i, C_j \) where \( \tau \) is the threshold value.

I) Rule 1: if \( (\varphi_{ci} \land \varphi_{cj}) < \tau \) and \( \varphi_{cn} < \tau \) then \( C_i \) and \( C_j \) are similar.

II) Rule 2: if \( (\varphi_{ci} \land \varphi_{cj}) < \tau \) and \( \varphi_{cn} > \tau \) then \( C_i \) and \( C_j \) are not similar.

III) Rule 3: if \( (\varphi_{ci} \land \varphi_{cj}) > \tau \) and \( \varphi_{cn} < \tau \) then \( C_i \) and \( C_j \) are not similar.

IV) Rule 4: if \( (\varphi_{ci} \land \varphi_{cj}) > \tau \) and \( \varphi_{cn} > \tau \) then \( C_i \) and \( C_j \) are not similar.

**Rule 1:** if \( (\varphi_{ci} \land \varphi_{cj}) < \tau \) and \( \varphi_{cn} < \tau \) then \( C_i \) and \( C_j \) is similar and if the similarity is represented here as \( S_{ij} \) then the following condition is valid:

\[ (\varphi_{ci} \land \varphi_{cj}) \land \varphi_{cn} \rightarrow S_{ij} \]

**Proof:**

Now in order to prove \( S_{ij} \) is valid for the considered rule the negation of similarity or \( \neg S_{ij} \) is unsatisfied. In order to prove the rules are valid to prove the similarity between \( C_i \) and \( C_j \), following are considered as an element of set \( S \):

\[ S = \{ \varphi_{ci}, \varphi_{cj}, \varphi_{cn}, (\varphi_{ci} \land \varphi_{cj}) \land \varphi_{cn} \rightarrow S_{ij} \} \]

The rules are now converted to the conjugate normal form as:

\[ S = \{ \varphi_{ci}, \varphi_{cj}, \varphi_{cn}, \neg \varphi_{ci} \land \neg \varphi_{cj} \lor \neg \varphi_{cn} \lor S_{ij} \} \]

Now consider, \( S' = S \cup G' \), where \( G' = \neg S_{ij} \) and hence the \( S' \) is as follows:

\[ S' = \{ \varphi_{ci}, \varphi_{cj}, \varphi_{cn}, \neg \varphi_{ci} \land \neg \varphi_{cj} \lor \neg \varphi_{cn} \lor S_{ij}, \neg S_{ij} \} \]

So, the five elements of set \( S' \) are as follows: 1. \( \varphi_{ci} \land \varphi_{cj} \), 2. \( \neg \varphi_{ci} \land \neg \varphi_{cj} \lor \neg \varphi_{cn} \lor S_{ij}, \neg S_{ij} \)

From (1) and (4): \( \varphi_{ci} \land \neg \varphi_{ci} \land \neg \varphi_{cj} \lor \neg \varphi_{cn} \lor S_{ij} \equiv \neg \varphi_{cj} \lor \neg \varphi_{cn} \lor S_{ij} \quad (6) \)

From (2) and (6): \( \varphi_{cj} \lor \neg \varphi_{cj} \lor \neg \varphi_{cn} \lor S_{ij} \equiv \neg \varphi_{cn} \lor S_{ij} \quad (7) \)

From (3) and (7): \( \varphi_{cn} \lor \neg \varphi_{cn} \lor S_{ij} \equiv S_{ij} \quad (8) \)

From (5) and (8): \( \neg S_{ij} \lor S_{ij} \equiv [\Phi] \)

Hence it is proved that the: \( S \models G \)

This implies that the uses of the above mentioned rule can be valid to prove that the \( C_i \) and \( C_j \) clusters are similar.
Rule 2: if \((\varphi_{ci} \lor \varphi_{cj}) < \tau \) and \(\varphi_{cn} > \tau\) then \(C_i\) and \(C_j\) are not similar and if the similarity is represented here as \(S_{ij}\) then the following condition is valid:

\[
(\varphi_{ci} \lor \varphi_{cj}) \land \lnot \varphi_{cn} \rightarrow S_{ij}
\]

So, \(S_{ij}\) which represents the dissimilarity between two clusters, is valid for the above mentioned rule.

Proof:

Now in order to prove whether \(S_{ij}\) is valid for the considered rule the negation of similarity or \(\lnot S_{ij}\) is unsatisfied. In order to prove the rules are valid and to prove the similarity between \(C_i\) and \(C_j\), following are considered as an element of set \(S\):

\[
S = \{\varphi_{ci}, \varphi_{cj}, \varphi_{cn}, (\varphi_{ci} \lor \varphi_{cj}) \land \lnot \varphi_{cn} \rightarrow S_{ij}\}
\]

The rules are now converted to the conjunctive normal form as:

\[
S = \{\varphi_{ci}, \varphi_{cj}, \varphi_{cn}, \lnot \varphi_{ci} \land \lnot \varphi_{cj} \lor \varphi_{cn} \lor S_{ij}\}
\]

Now consider,

\[
S' = S \cup G', \text{where } G' = \lnot S_{ij}\text{ and hence the } S' \text{ is as follows:}
\]

\[
S' = \{\varphi_{ci}, \varphi_{cj}, \varphi_{cn}, \lnot \varphi_{ci} \land \lnot \varphi_{cj} \lor \varphi_{cn} \lor S_{ij}, \lnot S_{ij}\}
\]

So, the five elements of set \(S'\) are as follows: 1. \(\varphi_{ci}\), 2. \(\varphi_{cj}\), 3. \(\varphi_{cn}\), 4. \(\lnot \varphi_{ci} \land \lnot \varphi_{cj} \lor \varphi_{cn} \lor S_{ij}\), 5. \(\lnot S_{ij}\)

From (1) and (4):

\[
\varphi_{ci} \land \lnot \varphi_{ci} \land \lnot \varphi_{cj} \lor \varphi_{cn} \lor S_{ij} \equiv \lnot \varphi_{cj} \lor \lnot \varphi_{cn} \lor S_{ij} \quad (6)
\]

From (2) and (6):

\[
\varphi_{cj} \land \lnot \varphi_{cj} \lor \varphi_{cn} \lor S_{ij} \equiv \lnot \varphi_{cn} \lor S_{ij} \quad (7)
\]

From (3) and (7):

\[
\varphi_{cn} \land \varphi_{cn} \lor S_{ij} \equiv \varphi_{cn} \lor S_{ij} \quad (8)
\]

From (5) and (8):

\[
\lnot S_{ij} \land \varphi_{cn} \lor S_{ij} \equiv [\varphi_{cn}]
\]

Hence it is proved that the: \(S \not\equiv G\)

This implies that the uses of the above mentioned rule can be valid if \(C_i\) and \(C_j\) clusters are not similar.

The validity of other rules can be proved in the similar way which is also implemented to find non similarity between two clusters. Hence it is possible to merge the two clusters on the basis of considered rules.

4.3.4. Evaluation:

The evaluation of the proposed method is of two folds. In the first fold the dataset and clustering accuracy calculation methods are described. In the second fold the performance of the proposed method is compared with four standard clustering algorithms. The data dimensionality effect is also studied by comparing the proposed method with other clustering algorithms.
The effectiveness of the proposed method is compared with four standard algorithms over nine artificial and benchmark datasets. The algorithms that are considered here for comparison are exception-maximization algorithm (EM) [20], DBSCAN algorithm [43], hierarchical clustering algorithm and k-means clustering algorithm.

The datasets used here for the clustering purpose are divided into two categories. Four synthetic data are used here and they are generated with different seed values. Those synthetic datasets are created with the different dimensions and classes [44,45]. The real time benchmark datasets are also used for comparison purpose. The details of the datasets are given in Table 4.1.

The clustering result are evaluated using the following accuracy measure

\[
A = \frac{\sum_{i=1}^{m} d_i}{n}
\]  

\(\text{(19)}\)

**Table 4.1. Dataset description**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of Instance</th>
<th>Data Dimensions</th>
<th>No. of Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic_1</td>
<td>633</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>synthetic_2</td>
<td>566</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Synthetic_3</td>
<td>1231</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Synthetic_4</td>
<td>1218</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Diabetes</td>
<td>768</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Sonar</td>
<td>208</td>
<td>60</td>
<td>2</td>
</tr>
</tbody>
</table>

where A, is the accuracy and \(d_i\) is the number of common data pattern between \(i^{th}\) cluster and its corresponding true cluster, and n is the number of data points. This accuracy measure based on Eqn.20 is also known as Huang’s accuracy measure [61]. According to the above mentioned accuracy measure the value of A must lie within 0 and 1, \(0 \leq A \leq 1\). The higher value of clustering accuracy is indicated by the nearest value of 1. Now, the overall accuracy of the clustering method for a single iteration is calculated.
as follows:

\[ \text{Clustering Accuracy} = \frac{\sum_{i=1}^{n} A}{n} \times 100 \quad (20) \]

In this experiment EM, DBSCAN, k-means and Hierarchical algorithm run for 50 iterations and for the EM, k-means and hierarchical algorithm, true number of cluster centers are provided for each dataset. Other parameter details are given in Table 4.2. The details of the real life dataset are as follows:

Data Set 1: The wine dataset is also multivariate. This data set consists of 178 numbers of instances and 13 numbers of attributes are found. The chemical composition is measured for three types of wine and used here as attributes. The wine data set consists of three classes of wine [46].

Data Set 2: The diabetes dataset consists of more than 768 instances and each consists of 8 numbers of attributes. The data contents are classified by one or two where one represents the case of tested positive for diabetes and two for test negative [47].

Table 4.2. Parameter Details

<table>
<thead>
<tr>
<th>K-means Clustering</th>
<th>EM Clustering</th>
<th>Hierarchical Clustering</th>
<th>DBSCAN Algorithm</th>
<th>Bat Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of iterations=50</td>
<td>no. of iterations=50; Minimum standard deviation = 1e-06; Seed value = 100</td>
<td>no. of iterations=50; seed value = 10</td>
<td>no. of iterations=50;</td>
<td>no. of iterations=100; Population size=25; ( \tau = 0.005; ) ( \alpha = 0.55 )</td>
</tr>
<tr>
<td>seed Value=10</td>
<td></td>
<td></td>
<td>C1=0.9; C2=6</td>
<td></td>
</tr>
</tbody>
</table>

Data Set 3: This ionosphere dataset consisting of 351 number of instances and the 34 number of attributes constitute one data. The data set consists of "good" and "bad" class and is represented using 1 and 0. The good one is represented by the existents of some kind of structure and in-case of bad class data are passed through the ionosphere [48].

Data set 4: The glass consists of 214 numbers of instances and each of them is created with 10 numbers of attributes. The chemical compositions of various types of glasses
are considered. The data set consists of 7 numbers of classes [49].

Dataset 5: The sonar dataset consists of 208 numbers of patterns and each of it is a result of transmitted sonar signal which is obtained at different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock. Each instance consists of 60 attributes representing the energy within a particular frequency band, integrated over a certain period of time [50].

The clustering algorithms are developed in the python programming language. The numerical analysis is performed in the algorithm with the "numpy" module and "matplotlib" module is used to plot the cluster. The experiment is carried out on an Intel Pentium 4 computer system which is configured with core2duo 1.66 GHz Intel processor, 2 gb of RAM and 500gb of hard disk space.

In Table.4.3 the performance of the proposed clustering method is shown in the form of accuracy measurement which is calculated in 50 numbers of iterations. Fig.4.1 depicts the results of the proposed clustering algorithm in 2d or 3d form.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>EM</th>
<th>DBSCAN</th>
<th>Hierarchical Clustering</th>
<th>K-means Clustering</th>
<th>BAT Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic_1</td>
<td>71.22</td>
<td>71.9</td>
<td>76.67</td>
<td>79.11</td>
<td>99.62</td>
</tr>
<tr>
<td>Synthetic_2</td>
<td>96.75</td>
<td>78.27</td>
<td>76.83</td>
<td>90.25</td>
<td>99.36</td>
</tr>
<tr>
<td>Synthetic_3</td>
<td>51.45</td>
<td>78.25</td>
<td>48.83</td>
<td>54.17</td>
<td>100.00</td>
</tr>
<tr>
<td>Synthetic_4</td>
<td>86.90</td>
<td>24.14</td>
<td>66.63</td>
<td>86.89</td>
<td>98.57</td>
</tr>
<tr>
<td>Wine</td>
<td>97.00</td>
<td>39.89</td>
<td>74.35</td>
<td>95.2</td>
<td>100.00</td>
</tr>
<tr>
<td>Diabetes</td>
<td>68.96</td>
<td>65.11</td>
<td>51.10</td>
<td>38.24</td>
<td>99.71</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>78.38</td>
<td>93.17</td>
<td>52.81</td>
<td>71.79</td>
<td>100.00</td>
</tr>
<tr>
<td>Glass</td>
<td>65.81</td>
<td>36.45</td>
<td>38.89</td>
<td>38.35</td>
<td>100</td>
</tr>
<tr>
<td>Sonar</td>
<td>59.66</td>
<td>97.00</td>
<td>53.54</td>
<td>59.97</td>
<td>93</td>
</tr>
</tbody>
</table>

The proposed method consists of parameters which are tuned empirically to enhance the clustering performance for different dataset. The values for the experiment over different dataset are decided after a long series of testing with a number of values. One threshold value used in the rule based merging, is presented in the Table.4.2 as τ. Another most important parameter in the proposed method is the population size of the
bat. In order to investigate the effect of the different population sizes over the clustering accuracy, the bat algorithm based clustering method is executed separately with 10,15,20,25,30,35,40,45,50 population size and rest of the parameter values are same as it is presented in Table 4.2. In Fig. 4.2 and Fig. 4.3 the effect clustering accuracy over different population sizes are analyzed for glass and wine dataset. The experiment shows that the algorithm performance is improved with 25 populations.

In Eqn.7 $\alpha$ parameter is used. This parameter is also treated as one of the convergence parameter and the effect of it is analyzed over ionosphere and sonar dataset which is represented here in Fig.4.4 and Fig. 4.5. The value of this parameter is decided after long range of testing over different dataset with different number of values and as the value of $\alpha$ is always less than one , the values for testing is always within the range of 0 and 1. Finally 0.01, 0.15,0.45,0.55,0.65,0.75,0.85 and 0.95 are decided for experimentation with different dataset and here experiment with these values over ionosphere and sonar datasets are shown. It is found from the experiment that the algorithm performance is improved if the value of $\alpha$ is set to 0.55 or 0.65. Another convergence coefficient involves in the rule based merging and it is denoted in the proposed method as $\tau$. This parameter effects the automatic cluster formation which directly effects the clustering accuracy. The effect of this parameter ($\tau$) on clustering accuracy is experimented over different dataset and here the experiment over ionosphere and synthetic_2 dataset are shown. The values of $\tau$ are randomly selected between 0 and 0.0001 for experimentation over different datasets and finally 0.001,0.005,0.01,0.09,0.1,0.5,0.9,0.95 are decided for experimentation with different dataset. The effects of experiment with the different values for ionosphere and synthetic_2 datasets are depicted in Fig. 4.6 and Fig. 4.7 respectively. It is observed from the experiment that if this parameter value is within the range of 0.001 to 0.005 then the algorithm performance is significantly improved. Now, at the time of experimentation with different values of parameter $\tau$, the value for other two parameter i.e. $\alpha$ and population size is fixed at 0.55 and 25. The result of these parameter settings is also fixed for ionosphere and synthetic_2 dataset (Fig. 4.6 and Fig. 4.7) and each of the experiment is set for 100 independent runs. So, the combination of these parameter settings (Table.4.2) is applied for all the dataset and the result of bat algorithm based clustering is presented in Fig. 4.1.
Fig. 4.1. Synthetic and real life benchmark dataset clustering using proposed method
Fig. 4.2. The effect on clustering accuracy over glass due to different population size

Fig. 4.3. The effect of clustering accuracy over wine dataset due to different population sizes
Fig. 4.4. The effect of different $\alpha$ parameters in clustering accuracies over Ionosphere dataset.

Fig. 4.5. The effect of different $\alpha$ parameters in clustering accuracies over Sonar dataset.
Fig. 4.6. Effect for different values of $\tau$ ($T$) over clustering of ionosphere dataset.

Fig. 4.7. Effect for different values of $\tau$ ($T$) over clustering of Synthetic_2 dataset.
4.4 Conclusion:
The presented work proposed a novel method of automatic clustering, known as bat algorithm based clustering technique. The proposed method provides almost accurate number of clusters over every synthetic and real life benchmark linear and non-linear datasets, presented here for the experimental purpose. The experimental results show that the proposed method produces an improved result which outperforms all the compared methods over the limited real life and synthetic datasets used here for the experiment. The proposed method incorporates a new rule based merging technique which removes the null clusters and also merge the relevant cluster on the basis of a defined threshold value. The effect of population size and other parameters are also analyzed, and an empirical way to choose the best combination of parameter values for bat algorithm based clustering method is provided, by means of a detail experiment over different sets of parameter values. The final selection of these parameter values are reported in Table.4.2. The future research should focus on the automatic and intelligent selection of the parameters of this algorithm, so that it can be applied to any environment.


4.5 Reference:


[26] Paterlini, S., Minerva, T,"Evolutionary Approaches for Cluster Analysis". In: 


[40] Airas, M., "Echolocation in bats", in Proceedings of Spatial sound perception and reproduction (the postgrad seminar course of HUT Acoustics Laboratory), April 2003.


[50] Sonar


[61] Swagatam Das, Ajith Abraham and Amit Konar, Automatic Clustering Using an