ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/95169, October 2016* ISSN (Online) : 0974-5645

An Efficient Clustering Approach using Hybrid Swarm Intelligence based Artificial Bee Colony- Firefly Algorithm

S. Karthikeyan* and E. J. Thomson Fredrik

Department of Computer Science, Karpagam University, Coimbatore - 641021, Tamil Nadu, India; s.karthics@gmail.com, thomson500@gmail.com

Abstract

Objectives: Extracting relevant information from large database is attaining huge significance. Clustering of relevant information from large database becomes difficult. The major objective of this work is to proposed novel clustering methods for solving clustering problem. **Methods/Statistical Analysis:** This proposed work introduces possibility of a novel approach Hybrid Artificial Bee Colony-Firefly Algorithm (HABC-FA) for clustering to solve the clustering problem in the benchmark datasets like Fisher's iris dataset. Here this FA incorporates its genome behavior of fireflies to accomplish the optimal clustering solution with ABC. The performance of this novel algorithm Hybrid ABC-FA is then compared with existing clustering algorithms like the ABC and hybrid Particle Swarm Artificial Bee Colony (PSABC) with regard to different statistical criteria making use three different types of benchmark datasets. **Findings:** The experimentation results prove that the proposed scheme performs better than the existing Swarm Intelligence (SI) based algorithms like ABC and PSABC in terms of speed and success rate and the proposed HABC-FA algorithm performance evaluates by using clustering parameters like recall, precision and F-measure. **Application/Improvements**: HABC-FA is proposed for the purpose of solving the clustering problem in the benchmark datasets like Fisher's iris dataset

Keywords: Artificial Bee Colony Algorithm (ABC), ABC-Particle Swarm Optimisation (PSO) (PSABC), Clustering, Hybrid Artificial Bee Colony with Firefly Algorithm (HABC-FA), Swarm Intelligence (SI)

1. Introduction

In general, clustering algorithms are divided into two important classes of algorithms known as the supervised and unsupervised. In the case of supervised clustering, the learning algorithm is provided with a guide that represents the goal class to which a data vector has to belong. Cluster partition is supposed to have the properties such as: (1) homogeneity inside the clusters, that is, data that fit to the same cluster must possess more similarity, and (2) heterogeneity among clusters, that is, data that fit to various clusters must be as diverse as possible. Various unsupervised clustering algorithms that are developed like the K-Means, ISODATA, learning vector quantizes (LVQ). Particularly K-means is a popular successful clustering scheme that is a center based, quick and simple algorithm. But, due to the K-means algorithm converging to the closest local optimum from its initial position, the algorithm's success is greatly dependent over the initial state with respect to the center of the cluster. Various methods like the statistics, expectation maximization algorithms, graph theory, evolutionary computing, artificial neural networks and swarm intelligence algorithms are employed forgetting over the local optima issue on clustering.

In the past decade, more number of random, optimization algorithms that are population-based has been used for clustering issues. The rising body of Swarm Intelligence (SI), which is metaheuristic algorithms include Ant Colony Algorithm (ACO)[1]Artificial Bee Colony (ABC)[2,3] Particle Swarm Optimization (PSO)[4], Firefly Algorithm (FA)[5,6], Glow worm Swarm

* *Author for correspondence*

Optimization (GSO)[7], Bacterial Foraging Optimization (BFO)[8], Biogeography-based optimization (BBO)[9], Cuckoo Search (CS)[10], Bat Algorithm (BA)[11], and flower pollination algorithm techniques[12] are used to overcome clustering problems recently. The swarm intelligence clustering schemes have benefits in several features, like no necessity of earlier information, and self-organization. Nonetheless, the amount of resultant cluster is extremely high often and moreover the convergence is extremely slow owing to the ant's in effective behaviours: arbitrarily taking and dropping items. Clustering takes part a substantial role in various fields inclusive of engineering (electrical, mechanical, industrial engineering), computer sciences, earth sciences, life and medical sciences, social sciences, and economics. In particular, the Artificial Bee Colony (ABC) algorithm is a considerably novel Swarm Intelligence scheme for clustering[13].

In case of ABC, the each one of the bee is considered as the iris dataset samples with their attributes. A bee that is waiting in the dance space for the purpose of making a judgment to group iris dataset samples points to choose optimal cluster iris dataset points is called onlooker and one which is moving to the nearest iris cluster data point food source that is passed by it already is called employed bee. The other category of bee is scout bee which conducts a stochastic hunt for noticing new sources. The location of a present iris dataset points cluster source points out a probable solution for the clustering issue and the nectar quantity of an iris cluster data point's food source is associated with the quality (fitness) is considered as the mean value of the iris dataset attributes of the corresponding solution, calculated by this work as using Firefly Algorithm. While ABC[14] using in the clustering process, it generates a good quality clusters in comparison against the rest of the population dependent algorithms though with lesser energy efficacy, consistency and generally with slow convergence speed[14].

This proposed work introduces and describes a resolution for the purpose of solving the above mentioned clustering problem, using one of the more efficient metaheuristic nature-inspired algorithms called Firefly[6]. The ABC and FA schemes are completely dependent on a specific successful mechanism of a biological phenomenon of Mother Nature, with the aim of accomplishing clustering solution, for instance, the family of honey-bee algorithms, in which the finding of an optimal solution is dependent on the foraging and the storing of undetermined amount of nectar. The Firefly algorithm belongs to the Swarm Intelligence (SI) based algorithm. The fireflies in the FA is used their behaviour function such as flashing light intensity for finding best solution in clustering process, which assists the swarm of fireflies for the purpose of moving to brighter and more attractive locations with the aim of obtaining efficient optimal solutions. For this behaviour of the fireflies, this algorithm combined with another SI based ABC algorithm for finding best solutions in the Clustering problem such as high energy efficiency, more reliability and high convergence speed. Hence in this work used highly more efficient Hybrid ABC - FA algorithm for resolving the problem of clustering, in this work that has been tested on various benchmarked data sets obtained from the UCI Machine Learning Repository.

2. Related Work

Particle Swarm Optimization (PSO)[15] is a technique for the purpose of clustering sophisticated and linearly inseparable datasets, with no previous knowledge regarding the amount of natural occurring clusters. This newly introduced technique is dependent over on an enhanced alternate of the PSO scheme. Moreover, it uses a kernel-induced similarity measure rather than the traditional sum-of-squares distance. Usage of the kernel function renders it feasible for clustering the data which is linearly inseparable in the actual input space into homogeneous clusters present in a high-dimensional feature space that is transformed. Computerized simulations have been carried out with the help of a test bench consisting of 5 artificial and 3 original life datasets for the purpose of comparing the performance of the newly introduced technique with less number of modern clustering schemes. The results reflected that the superior behavior of the newly introduced algorithm in accordance with accuracy, convergence speed and reliability.

New Artificial Bee Colony (NABC)[16] algorithm completely transforms the search pattern of both employed and onlooker bees. A solution pool is built through the process of storing some best solutions of the existing swarm. With this, new candidate solutions are produced through the process of searching the neighbourhood of solutions randomly chosen from the solution pool. Experimentations are done on a set of 12 benchmark functions. Results confirm that this scheme is considerably better or at least similar to the original ABC and seven other stochastic schemes.

pyG Cluster[17] which is a clustering algorithm that is focused upon noise injection for the next subsequent cluster validation. The reproducibility of a huge amount of clusters got with agglomerative hierarchical clustering is evaluated. Moreover, a variety of diverse distance-linkage combinations are assessed. At last, largely reproducible clusters are then meta-clustered into communities.

Genetic Algorithm (GA)[18] for clustering process with the purpose of testing the functioning of the algorithm proposed, it was coded and made to function over a stochastic data set. This work saw that in common applications, the algorithm's performance was equal to that of the k-Means clustering. This work made transformations to the fitness function and proved that the behavior of the algorithm was satisfying the expectations and produced much better clusters compared to the k-Means clustering algorithm, which was the desired result.

Unsupervised reliable clustering algorithm[19] is proposed which can discover dense areas in feature space successfully and decide over their quantity. The major complication of clustering is then modified into a multimodal function optimization issue inside the area of genetic niching. The identification of the niche peaks, constituting the last cluster centers, is done on the basis of Deterministic Crowding (DC). The issue of crossover communications in DC is removed through limiting the mating to only the members belonging to the same niche. At last, the right amount of niche highest or the cluster centres is subsequently extracted from the end population. Genetic optimization renders this method which is much less susceptible to suboptimal solutions rather than the rest of the objective function dependent techniques, and provides it freedom from requiring an analytical derivation corresponding to the prototypes. Consequently, this technique can deal with a broad array of the usual subjective, also the non-metric dissimilarities, and hence is helpful in several applications like Web and data mining. Moreover, the usage of reliable weights offers it to have a reduced amount of sensitivity to the existence of noise compared to the majority of the conventional unsupervised clustering approaches.

3. The Problem Statement of Clustering Process

Clustering is the procedure of dividing a provided dataset of $\mathbf{n}$ data points into K groups or clusters in accordance with the certain similarity (distance) metric between the data points of the iris datasets. Let $\mathbf{Ir} = \left\{\overrightarrow{\mathrm{Ir}_1}, \overrightarrow{\mathrm{Ir}_2}, \ldots \ldots \overrightarrow{\mathrm{Ir}_r}\right\}$ represents a set of N data objects to be clustered and every data object in the iris dataset is represented as $\overrightarrow{\mathrm{Ir}_l} = \{\mathrm{Ir}_{i1}, \mathrm{Ir}_{i2}, \ldots, \mathrm{Ir}_{id}\}$ where $\mathbf{Ir}_{id}$ indicates a feature value of data object $\mathbf{Ir}$ on dimension $\mathbf{d}$ .The major objective of the clustering is to select over a set which includes K partitions $\mathbf{C} = \{\mathbf{C}_i, \ldots, \mathbf{C}_k\}$,subsequently:

$$\mathbf{C}_i \neq \emptyset, \mathbf{for\ i} = 1, \ldots, \mathbf{k}, \tag{1}$$

$$\mathbf{C}_i \cap \mathbf{C}_j \equiv \emptyset, \mathbf{for\ i} = 1, \ldots, \mathbf{k}, \mathbf{j} = 1 \ldots, \mathbf{k}, \mathbf{and\ i} \neq \mathbf{j}$$

$$\mathbf{and}\ \cup_{k-1}^{K} \mathbf{C}_k = \mathbf{0} \tag{2}$$

Every cluster is denoted by a cluster center, hence refers to a set possessing the cluster centers to which the data objects are allocated. This work will employ the Euclidean metric in the form of the distance metric between the iris dataset's data point in cluster.

$$\mathbf{I} = \left\{\overrightarrow{\mathrm{I}_1}, \overrightarrow{\mathrm{I}_2}, \ldots \ldots \overrightarrow{\mathrm{I}_N}\right\} \overrightarrow{\mathrm{Z}_l} = \frac{1}{\mathbf{n}_i} \sum_{\mathrm{O}_i \in \mathrm{c}_i} \overrightarrow{\mathrm{O}_l} \tag{3}$$

Where $\mathbf{n}$ indicates the number of data objects in cluster.

4. Proposed Methodology

Artificial Bee Colony (ABC)[20] scheme for the purpose of optimization of numerical issues. This scheme does the simulation of the smart foraging behavior of the honey bee swarms. It is extremely uncomplicated, strong and population dependent stochastic optimization algorithm. The performance obtained from the ABC algorithm is enhanced with SI based metaheuristic algorithms like FA on clustering problem problems. The performance of HABC-FA algorithm on benchmark dataset tested on clustering problems. In Cluster analysis of using the proposed methodology HABC-FA, that groups Iris Data Sets' data objects into clusters like objects that belong to the same cluster are similar which is dependent on the features, while those which belong to diverse ones are dissimilar by classifying the iris datasets based on their classes such as Iris Setosa, Iris Versicolour, Iris Virginica.

4.1 Clustering Analysis

The clustering problem cannot be solved by using a one-step process. In one of the benchmark iris datasets is

used in the proposed method for the purpose of solving the clustering complication. The Figure 1 shows that the proposed cluster analysis processing strategies.

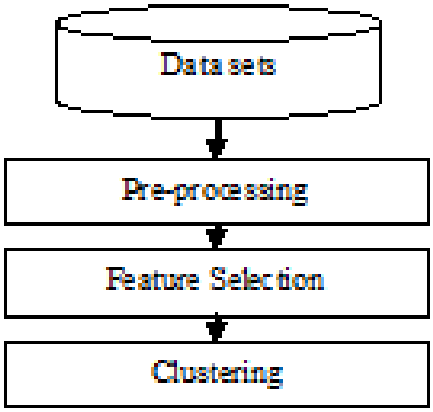

Figure 1. The block diagram of proposed methodology.

4.1.1 Datasets

The UCI repository Database Irish dataset is used in this work, which is a public domain online database which collects experimentally obtained benchmark datasets from various sources. This work focus on the benchmark datasets such as Iris Dataset used for solve the clustering problem with proposed algorithm HABC-FA. In this technical work, data objects of the Iris datasets are distinguished by their individual feature values for a set of attributes such as sepal length, sepal width, petal length and petal width.

4.1.2 Pre-Processing

Most of the clustering methods depend on various pre-processing techniques to achieve optimal quality and performance in the clustering problem. The pre-processing techniques include removal of redundant and irrelevant features from the datasets. Irrelevant features, together with redundant features, rigorously disturb the accuracy of the clustering process in learning machines datasets. As a result, feature subset selection must be capable of identifying and eliminate as much of the irrelevant and redundant information as possible. Furthermore, a good feature subsets contain features extremely correlated with the class, however uncorrelated with (not predictive of) each other. The irrelevant feature removal is simple when the right relevance measure is determined or selected, however the redundant feature elimination is a bit of sophisticated.

Here this work used ICA for removal of irrelevant features from the dataset. In clustering applications, it is believed that the components that are independent expose something remarkable of a multi-dimensional data set.

4.1.3 Independent Component Analysis

This work used ICA model[21] $m = Za$ of independent features a of iris dataset and mixing matrix Z. However, estimated practically, is actually the demixing matrix Q for $a = Qm$, where Q stands for a (pseudo) inverse of Z. The algorithm is executed L times on $dataM = [m_1, m_2 \ldots m_N]$ comprising of N samples having k-dimensional vectors. The estimations with respect to the demixing matrices $\widehat{Q_i}$ from every run $i = 1, 2, \ldots, L$ are gathered in a single matrix $\hat{Q} = \left[\widehat{Q_1^T}, Q_2^T \ldots, Q_L^T\right]^T$. In case if n_i independent features are estimated on each round, and obtain $K = \sum_i n_i$ estimates, and subsequently the size of $\hat{Q}$ will be $S \times s$. W

The normal evaluation similarity seen among the estimated independent features is the complete value of their individual mutual correlation coefficients $Cr_{ij}, i, j = 1, \ldots, S$. Straightforward computations reveal that it can be obtained as the components of R = Q ΣQ^T where Σ indicates the covariance matrix for M.

The last similarity feature matrix has then components, $\sigma_1 ij = |Cr_{\downarrow} ij|$. W

Later, for clustering techniques and validity indices which anticipate dissimilarities in their standardized form. Then this work transforms the similarity feature matrix into a dissimilarity matrix having elements ds_{ij}. A much easier method to perform this transformation is apparently[22] :$ds_{ij} = 1 | \sigma_{ij}$. After removal of irrelevant features and generated the similarity matrix used by ICA, the feature selection stage involves in the iris dataset in the proposed work.

4.1.4 Feature Selection

The process of feature selection includes the procedure for identifying a suitable subset of the most valuable features in the iris dataset used in this work, which gives well-matched clustering results as the complete iris data set of features.

Consider **G** is feature subset of **F** datasets and **fG**

represents the value data point vector of G feature subset. Normally, the major objective of feature selection is the selection of a minimum feature subset of $\mathbf{G}$, such that $\mathbf{P(C \mid G = fG)}$ is equivalent or as near as possible to $\mathbf{P(C \mid F = f)}$, where $\mathbf{P(C \mid G = fG)}$ is the probability distribution of different classes of dataset provided the feature values in $\mathbf{G}$ and $\mathbf{P(C \mid F = f)}$ indicates the actual distribution provided the feature values in **feature subset of F**. This is known as minimum subset of an optimal subset. In this proposed work used Genetic Algorithm for find an optimal feature subset in feature selection process.

4.1.5 Feature Selection using GA

Genetic algorithm (GA) is well-known for their capability of effectively searching vast spaces about which little knowledge is available. While GA is relatively thoughtless to unnecessary features in the dataset, they are considered to be an interesting option for the basis of a further strong feature selection technique for the purpose of enhancing the performance of clustering process. In this section this proposed work describes GA for feature selection in detail.

4.1.6 Genetic Algorithm

GA considered being a process of inductive learning technique, which is adaptive search strategy, has showed a considerable enhancement over a range of random and local search techniques[23]. This is achieved through their capability of exploiting the collecting information regarding an initial unknown search space for the purpose of biasing the next successive search into hopeful subspaces. In the meantime, GA is originally a domain free search method, they are suitable for several applications in which the domain information and theory is hard or not feasible to be provided. The important challenges in using the GA to any issue are the selection of an ideal demonstration and also a sufficient evaluation function.

For the description in detail regarding both these challenges for the complication concerned with feature selection[24]. In the problem of feature selection the chief concentration lies in the representation of the space of every probably subsets of the feature set given. Subsequently, the simplest method of the representation corresponds to the binary representation, in which, every feature present in the candidate feature set is regarded as a binary gene and every individual comprises of fixed-length binary string that represents certain subset of the feature set given. An individual having length l associates with a $l -$ dimensional binary feature vector X, in which every bit indicates the removal or insertion of the feature associated. Subsequently, $x_i = \mathbf{0}$ indicates the removal and $x_i = \mathbf{1}$ represents the insertion of the ith feature. The selected best feature subsets are used for finding the best clustering result using proposed HABC-FA algorithm in clustering process[25], the proposed clustering process described in detail as follows.

4.1.7 Clustering Process

After finding the best feature subset for the clustering process the proposed clustering Strategy that involves the careful choice of clustering Algorithm for solving the clustering problem. Here this proposed work used swarm intelligence based HABC-FA for finding optimal solution in the clustering problem. The detail description of the ABC and Firefly Algorithm described as follows.

4.1.8 ABC Algorithm

ABC algorithm is a novel population-based metaheuristic technique,. It has found its application in several complicated issues. This scheme simulates the intelligent foraging behaviour of honey bee swarms. This scheme tends to be extremely uncomplicated and reliable.

In general the colony of artificial bees in the ABC is divided into three groups: employed bees, onlookers, and scouts. Employed bees correspond to a certain food source which they are exploiting at present or which they are "employed" at. They transmit the details regarding this specific source with them and then convey this information with the onlookers. Onlooker bees are those bees which are waiting for the information to be shared through the employed bees on the dance space in the hive regarding their food sources, and thereafter they decide about choosing a food source. A bee that conducts anarbitrary search is known as a scout. In the case of the ABC algorithm, the initial half of the colony includes the employed artificial bees and the second half includes the onlookers. For every food source, there is one solitary employed bee. Else, the amount of employed bees equals to the amount of food sources existing around the hive. The employed bee whose food source has been exhausted by the bees then turns out to be a scout. The place of a food source specifies a probable solution for the complication of optimization and the quantity of nectar in a food source

is related with the quality fitness of the particular solution indicated by that food source. Onlookers are positioned on the food sources with the assistance of a probability-based selection procedure. When the amount of nectar in a food source is increased, the probability value of the food source with which it is selected by the onlookers also increases.

4.1.9 Firefly Algorithm

FA[26] gets its inspiration by means of the biochemical and social concepts of actual fireflies. Actual fireflies generate a short-timed and rhythmic flash which assists them in having their mating partners attracted towards them and more overacts as a precautionary warning mechanism. Here the Firefly Algorithm formulates this real fireflies flashing behavior and used this behavior in the onlooker phase bee's genome with the same data set objective function of the clustering issue which is to be optimised with HABC algorithm.

Consider the cluster classified issue in which the task is about minimizing the iris dataset objective function $\mathbf{f_{ir}(x)}$ for $\mathbf{x \in S \subset \mathfrak{R}^n}$, i.e., determine $\mathbf{x^*}$ using,

$$\mathbf{f_{ir}(x^*) = \min_{x \epsilon X} f_{ir}(x)} \quad (4)$$

The swarm of $\mathbf{n}$ fireflies is denoted data points in the cluster to resolve the above said problem in an iterative manner and $\mathbf{x_i}$ represents a best distance value of the data point in the group in the clustering problem resolution for a firefly $\mathbf{i}$ during the iteration $\mathbf{t}$, in which $\mathbf{f(x_i)}$ represents its fitness distance value of data point. The fitness of every data point is decided by the landscape corresponding to the data set objective function. Moreover, this fitness value decides over the attractiveness of each $\mathbf{i^{th}}$ member existing in the data point and indicated its attractiveness as light intensity $\mathbf{I_i}$.In the beginning, the entire data points are seen to be dislocated in S (either arbitrarily or by using certain deterministic technique). Each data points discovers its matting partner such as same features data points in the cluster on the basis of the attractiveness based on the feature of other data point in the cluster space and travels towards that neighbourhood dataset for the purpose of improving its fitness of the distance of the data points in the category. The next subsequent position of data point $\mathbf{i}$ during iteration $\mathbf{t+1}$ is decided according to equation (8) which takes two factors first into consideration, which are the attractiveness of the other swarm data points with greater light intensity, i.e $\mathbf{., I_j > I_i, \forall j, j = 1, \dots \dots n, j \neq i}$, which is changing across distance and seconds a pre-determined random step vector $\mathbf{u_i}$.

$$\mathbf{x_i\ (t\ +1) = \ (1\ -\ \beta) x_i\ (t) + \ \beta x_j\ (t) + \ \mu_i} \quad (5)$$

where β is the attractiveness such as best fitness distance value of data point $\mathbf{j}$. It shows how robust it fitness value of data point $\mathbf{i}$ in the group and calculate using equation (9).

$$\beta \ = \ \beta \mathbf{e^{\gamma r_{ij}}} \quad (6)$$

where $\mathbf{r_{ij} = d(x_i, x_j)}$,a Euclidean distance among two data points $\mathbf{i}$ and $\mathbf{j}$ .On the whole,$\beta_0 \in [0, 1]$,defines the fitness value o distance during $\mathbf{r = 0}$, i.e., when two data points are observed at one point of the search space S. The value of $\gamma \in [0, 10]$ plays a major part in deciding the variation seen in the fitness value as the distance is increased from the communicated data points. It is fundamentally the light absorption coefficient and normally$\gamma \in [0, 10]$ could be suggested[26].

Based on the fundamental behaviour of these algorithms, this proposed work used HABC-FA algorithm for the purpose of resolving the clustering complications and the main steps of the proposed algorithm are given as follows:

The proposed work using HABC-FA for clustering process, In this proposed hybrid algorithm, at first the initialization stage in the ABC algorithm produces a stochastically distributed initial food source like the data point positions in the cluster of SN solutions, where SN indicates the size of data points i.e. employed bees or onlooker bees. Every solution $\mathbf{\square ir \square^{\uparrow} I, (i = 1, 2, \dots, SN}$ is a D-dimensional data vector. Here, D refers to the amount of optimization parameters. Subsequently, each nectar amount $\mathbf{fit_i}$ is assessed.

During the employed bees' stage, each employed bee notices a new food source, for example, data point $\mathbf{V_i}$ in the neighbourhood of its present data point source $\mathbf{ir_i}$ **in clustering process,** the new data point cluster is computed making use of the expression that follows:

$$\mathbf{v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj})} \quad (7)$$

where $\mathbf{k \in 1, 2, \dots, SN}$ and $\mathbf{j \in 1, 2, \dots, D}$ are arbitrarily chosen indexes, and $\mathbf{k \neq i.}\ \phi_{ij}$ represents a random number between [−1, 1]. Subsequently, employed bee compares the new one with the current clustering solution and remembers the better one through greedy selection mechanism in the employee bee phase.

In the onlooker bees' phase, every onlooker selects a data point having a probability that corresponds to the

nectar quantity fitness of a data point distance value, calculated by using FA, which is described in the below section. The best fitness distance value is shared through the employed bees. Probability is determined with the assistance of the following equation:

$$\mathbf{p^i} = \frac{\mathbf{fit_i}}{\sum_{n=1}^{SN} \mathbf{fit_i}} \quad (8)$$

During the scout bee phase, just in case if a data point cannot be improved with the assistance of a fixed number of cycles, known as the "limit", it is removed from the data source, subsequently the employed bee belonging to that data point becomes scout. The scout bee subsequently discovers a new random data point location for increasing the clustering process, which makes use of the following equation:

(9)

where $\mathbf{x_i min^{\uparrow} j}$ and $\mathbf{x^j_{max}}$ are the lower and upper limits of parameter $\mathbf{j}$, data point respectively. All these steps are repeated through a fixed number of cycles, called as the Maximum Cycle Number (MCN), or until a criterion is met. The hybrid technique, which is defined as the combination of more than one technique, here in this proposed work used an Artificial Bee Colony (ABC) with Firefly Algorithm (FA) namely known as Hybrid ABC-FA, this hybrid algorithm used to find the best distance value of data point for best clustering solution for benchmark dataset such as iris dataset used in this clustering process. The performance of this hybrid algorithm is evaluated with help of UCI machine learning repository benchmark iris datasets.

The proposed method using the HABC-FA algorithm in the clustering strategy, the aim of every single bee in ABC is consider as data points of the iris dataset which is used to generate the best clustering solution. From expression (4), it can be seen that the new cluster of the iris dataset is generated by a stochastic neighbourhood data point of current clustering position and a stochasticone single dimension of D-dimensional vector of the dataset. This will give rise to a complication which a single data point might have found a better dimension, though the fitness corresponding to the individual data object group of the clustering process in the iris dataset is calculated by making use of D-dimensional vector, therefore it is extremely possible that the individual group is not the best solution finally and then the good dimension of the attractiveness of the artificial swarm, which the individual group of data object has found, it will be eliminated.

For the purpose of generating a good solution vector, every one of the data point groups of the ABC algorithm must cooperate with the FA and the information from each of the data point s of the attractive swarm (i.e. fireflies) is necessary to be brought into use., which will give the best distance value between the data point of the benchmark iris dataset. Therefore, these works apply Improved Hybrid ABC with FA to resolve the clustering problem in the clustering approach. In the HABC-FA algorithm, for this work consider a super best solution vector, which is, $\mathbf{new_{best}}$ and its every constituent of D-dimensional is considered to be the finest in all data point groups. For $\mathbf{new_{best}}$: $(n_1, n_2 \ldots, n_D)$ n_1 corresponds

Algorithm 1. Important steps of the HABC-FA algorithm

Step 1 Cycle=1
Step 2 Begin the data point group positions ir_i i = 1, ...SN
Step 3 Assess the nectar amount(fitness fit_1) of group data sources and discover the best group data source which is the preliminary value of new_{best}
Step 4 **Repeat**
Step 5 For each data point € (1, 2, ...,D)
Step 6 **Employed Bees' Phase**
For every employed bee i = 1... SN
Calculate the data point in group f_{ir} [$newn_{best}$ (n_1,n_2...., ir_{ij} ...n_D)]
If data point in the cluster better than f_{ir}($newn_{best}$)better than f_{ir}($newn_{best}$)
Then $newn_{bes}$ is replaced by $newn_{bes}$
For employed bee i generate new group data source positions with the assistance of **(4)**
Compute the cluster value fit_1
Implement greedy selection mechanism
End For
Calculate the datapoint nbest $f_{\downarrow ir}$ [$newn_{\downarrow best}$ $n_{\downarrow}1$, $n_{\downarrow}$ 2... . ir ij... . , g D)]

If data point f_{ir} ($newn_{best}$) better than f_{ir} ($newn_{best}$)
Then $newn_{best}$ is replaced by $newn_{best}$
For employed bee^{i} create new group data source positions with the assistance of **(4)**
Compute the cluster value fit_1
Implement greedy selection mechanism
End For
Calculate the probability cluster values p_1 for the clustering problem solution.

Step 7 **Onlooker Bees' Phase**

For every onlooker bee i = 1, ...,SN
Select a group based on data point source based on p_1 and applying the FA for attractive fitness function for finding best distance value
Generate an initial group randomly of n fireflies within D dimensional search space x_{ik}, i = 1, 2, ...N and k = 1, 2, ...k. Atractiveness of the FA evaluate the best fitness distance value of the group $f(x_i)$ which is directly proportional to light intensity I_i, tis a best distance value among the data point of the cluster.
To use these attractiveness of FA in ABC,
Replace the j data point of the group $newn_{best}$ by using the j data point of FA bee i
Calculate the $f_{\downarrow ir}[(newn_{\downarrow}(best\ n_{\downarrow 1,}\ n_{\downarrow}2...x_{\downarrow}ij\ ...\ .\ ,\ g_{\downarrow D})])$
If data point $f_{ir}(newn_{best})$ better than
Then $newn_{best}$ is replaced by data point $newn_{best}$
For onlooker bee ^{i} generate new group data source positions viwith the assistance of **(4)**
Compute the value fit_1
End For
End For

Step 8 **Scout Bees' Phase**

If there presents an employed bee it turns out to be scout
Then substitute it with a new random group data source positions in clustering

Step 9 Remember the best solution for clustering problem attained so far
Step 10 Compare the best solution with data point new n_{best} and remember the better one.
Step 11 Cycle= cycle + 1.
Step 12 **Until** cycle =Maximum Cycle Number

to the i-th data point of the $\mathbf{new_{best}}$ .The fitness function in the onlooker bee stage in the HABC is indicated as $\mathbf{f_{ir}}$ in algorithm 1.

The newly introduced HABC-FA algorithm does the recast of the above said Firefly algorithm scheme to apply with improved HABC, to enhance the clustering algorithm's performance. Each firefly in the FA for finding the distance value between the data points **j** is indicated in **D** dimensions in which each dimension represents the centroid of cluster and subsequently moves its location for the purpose of achieving the data objective functions of the iris datasets. The HABC-FA Algorithm will permit a firefly to traverse to get the best distance value between the data point of the selected cluster for best clustering results in use of the iris datasets. The proposed clustering algorithm increases energy efficiency, consistency and it will give high convergence speed, because of the best feature selection and pre-processing steps of this clustering approach.In this clustering stage, in proposed work includes the combination of clustering results with classification, in order to draw best clustering results. A good subset of features can not only enhance the accuracy of classification, however also considerably lessen the time to derive rules. It is implemented particularly when the amount of dataset attributes in a particular datasets is extremely large. As clustering results can characterize the basis for distribution of the whole data sets, clustering is helpful to aid supervised classification of datasets based on their selection of features. Thus, clusters can be used to select useful features and subsequently to add to sample datasets to improve the performance of classification in clustering process. Thus the classification steps used for the purpose of increasing the convergence speed of the

clustering approach. The following section shows that the experimentation results of the new technique.

5. Experimental Results

5.1 Data Sources

The proposed scheme has been assessed with the assistance of three data sets from different knowledge fields to know the effectiveness of proposed Hybrid ABC-FA algorithm. The data sets are available from UCI Machine Learning Repository. Table 1 provides a description of the tested data sets, together with the number of instances, number of features for each data set.

The Iris data set includes 50 samples from each of three species of Iris. Four features were taken from each sample: the length and the width of the sepals and petals. The Thyroid dataset consist of 175 samples, together with three features such as euthyroid, hyperthyroidism, and hypothyroidism patients. The Wisconsin breast cancer dataset comprises of 459 samples, with six features such as clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, and mitoses. Based on the combination of datasets features, this proposed work developed an efficient clustering approach HABC-FA for solving the clustering problem and the performance of this proposed algorithm compared with algorithm like ABC and PSABC.

Table 1. The sample of Bench mark Datasets

Bench Mark Datasets	No. of Samples	No. of features of the data object in the dataset
Fisher's iris dataset	50	Sepal length, sepal width, petals length and width
Thyroid dataset	175	euthyroid, hyperthyroidism, and hypothyroidism patients
Wisconsin breast cancer dataset	459	clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, and mitoses.

5.2 Statistical Criteria used in the Clustering Process

The statistical measures that follow are exploited for evaluating the results and compared generated results with HABC-FA by various clustering algorithms[27].

5.3 Minimization of Outline within Criteria (OLW)

This criteria equation completely depends on data objective collective within groups' dimensional matrix W. The collective within data cluster object D- Dimensional matrix W is given as,

$$\sum_{k=1}^{K} W_k \tag{10}$$

where W_k points out the variance matrix of the data object allocated to cluster C_k, where $k = \{1, \ldots, K\}$.

$$W_k = \sum_{i=1}^{n_k} \left(\overrightarrow{o_l^k} - \overrightarrow{o^k}\right)\left(\overrightarrow{o_l^k} - \overrightarrow{o^k}\right)^T \tag{11}$$

where $\overrightarrow{o_l^k}$ indicates the i^{th} data object in cluster C_k and n_k refers to the amount of objects in cluster C_k. and $\overrightarrow{o^k} = \frac{\sum_{i=1}^{n_k} \overrightarrow{o_l^k}}{n_k}$ indicates the vector of the centroid for the particular cluster C_k. Where K is number of groups or clusters on the basis of certain similarity (distance) metric among the data points of the datasets. A set of N data objects has to be clustered in the process of clustering.

5.4 Maximization of Variance Ratio Criteria

This criterion is dependent on data collective inside group's D-dimensional matrix W and among the group D-dimensional matrixes B. The between dimensional matrix B is given as per equation (3)

$$B = \sum_{k=1}^{K} n_k \left(\overrightarrow{o^k} - \vec{o}\right)\left(\overrightarrow{o^k} - \vec{o}\right)^T \tag{12}$$

Where, $\overrightarrow{O} = \frac{\left(\sum_{i=1}^{N} \overrightarrow{o_l}\right)}{N}$

Therefore the variance of criteria is VAR defined as follows,

$$VAR = \frac{\left(\frac{(\text{trace}(B))}{(K-1)}\right)}{\left(\frac{(\text{trace}(W))}{(N-K)}\right)} \tag{13}$$

The measurement of the efficiency of the algorithms is done based on the criterion that follows:

Mean best fitness value of OLW, VAR as given in equations (11) and (13). Successive percentage (%) that attain the best known data objective function value over the no. of simulations. The benchmark datasets are taken into consideration for evaluating the performance of the algorithms.

5.5 Mean Best Fitness Value and Variance

The mean best fitness values and variance of criteria o of the clustering analysis corresponding to the benchmarked datasets are listed in the Table 2 and Table 3. The result indicates that the HABC-FA provides best solution for clustering in both of the parameters such as mean best fitness value and VAR.

The Figure 2 illustrate that the performance i.e. mean best fitness value from the clustering processes.

Table 3. VAR criterion of the clustering analysis of benchmark datasets fitness values of the algorithms using in the clustering analysis of the given datasets.

The Figure 3 illustrate that the performance i.e. variance of the algorithms using in the clustering analysis of the given benchmark datasets.

Table 2. The Mean best fitness value in the clustering analysis of benchmark datasets such as Fisher's iris, Thyroid and Wisconsin breast cancer

Datasets	ABC	PSABC	HABC-FA
	Mean best Fitness value of OLW	Mean best Fitness value of OLW	Mean best Fitness value of OLW
Fisher's iris dataset	68.25	69.15	68.67
Wisconsin breast cancer dataset	68.30	67.85	69.75
Thyroid dataset	68.58	67.65	69.89
Mean Best fitness Value of OLW	68.37	68.22	69.44

Table 3. VAR criterion of the clustering analysis of benchmark datasets

Datasets	ABC	PSABC	HABC-FA
	VAR	VAR	VAR
Fisher's iris dataset	39.89	42.95	44.79
Wisconsin breast cancer dataset	43.45	44.15	44.85
Thyroid dataset	43.40	44.21	44.91

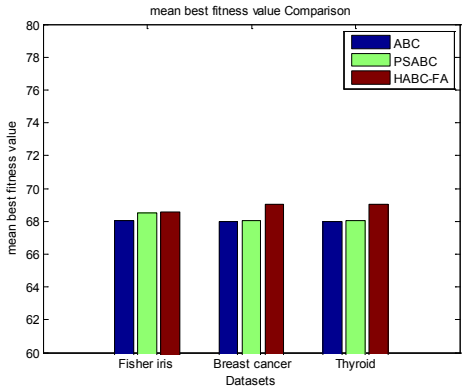

Figure 2. The comparison result of the mean performance of the proposed and existing clustering processes through Datasets..

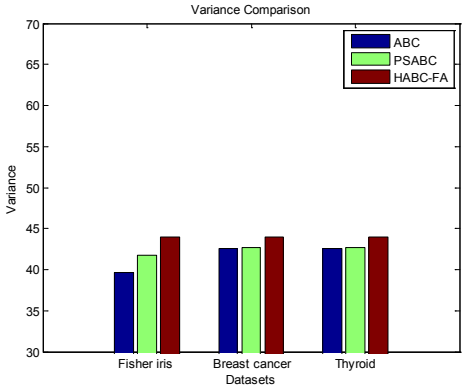

Figure 3. Variance of the clustering Process comparing different algorithms.

5.6 Success Percentage

The success rate which attain the extent best known objective function value based on percentage of number of runs (i.e., success %) is tabulated in Table 4 for benchmarked datasets. And Figure 4 illustrates the success rate of the newly introduced algorithm using in the datasets

The Figure 3 illustrates the percentage of number of runs (i.e., success rate in %) for benchmark datasets using the proposed HABC-FA and existing methods.

Table 4. Percentage of number of runs (i.e., success %) for benchmarked datasets

Datasets	ABC	PSABC	HABC-FA
Fisher's iris dataset	64	66	74
Wisconsin breast cancer dataset	81	90	92
Thyroid dataset	68	78	86

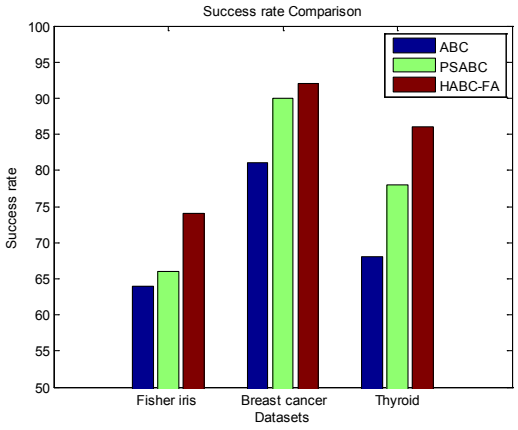

Figure 4. The success rate of the Proposed Algorithm and the Existing Algorithm.

5.7 Performance Analysis of proposed HABC-FA

Here this section focus on the evaluation of the proposed HABC-FA clustering validation measures including recall, precision and F-measure. This three are generally used clustering validation measures for clustering process.

Recall (RC)

The recall value is obtained with the no. of most relevant feature dataset of cluster and the total no. of relevant feature datasets in cluster.

$$\text{Recall(RC)} = \frac{\text{No. of most relevant feature dataset of cluster}}{\text{Total No. of relevant feature datasets in cluster}} \quad (14)$$

The Figure 5 shows the clustering performance assessment of the recall values of both proposed and existing methods for the given inputs datasets.

Precision (PC)

The precision value (PC) is computes with the total no. of relevant features in the datasets of cluster.

$$\text{PC} = \frac{\text{No. of relevant features dataset clustered}}{\text{Total no. of dataset features clustered}} \quad (15)$$

The Figure 6 depicts the performance comparison of the precision values of both proposed and existing methods.

F-measure:

The F-measure performance comparison for proposed and existing algorithms is calculated by combination of the precision and recall result values from the clustering

$$\text{F} - \text{measure} = \frac{2\ \text{RC.PC}}{\text{PC} + \text{RC}} \quad (16)$$

This work taken to treat each cluster results of a dataset, then compute the recall and precision of that cluster for all given datasets and F-measure is computed with the assistance of the equation (16),

The Figure 7 shows the f-measure comparison of the both proposed and existing methods.

The large F-measure value of proposed HABC-FA indicates that higher clustering quality of the proposed algorithm. From the observations from the Mean fitness value, VAR and success rate measure, altogether the three algorithms provide the best-known value to be within most no. of assessments and they have an increasing success rate to attain the optimal value for the dataset. The HABC-FA outperformed in many of all the cases. It is also observed that the convergence of the HABC-FA is extremely quick for the benchmark datasets like iris, Wisconsin breast cancer and thyroid. The observation obtained from the OLW for VAR metric is that the working of HABCA-FA is significant relating to the benchmark datasets. The HABC-FA algorithm has a greater probability for finding the necessary optimal value when compared to ABC and PSABC. Therefore, the convergence of the newly introduced HABC-FA for the Variance measure has a good speed when compared to other algorithms for many of the benchmark issues. The proposed HABC-FA clustering algorithm's performance is also measured using clustering process evaluation parameters like recall, precision and F-measure. However optimization based methods have provide better results for all applications some of the work are solves routing problem in Wireless Sensor Networks (WSNs)[28], ABC Based PID Controller[29] for Nonlinear Control Systems, Network Lifetime enhancement using PSO based Apriori[30] and Gravity Dam using PSO algorithm[31].

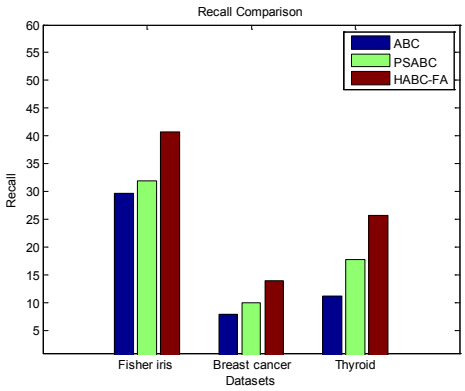

Figure 5. Recall comparison of proposed method Vs. existing method.

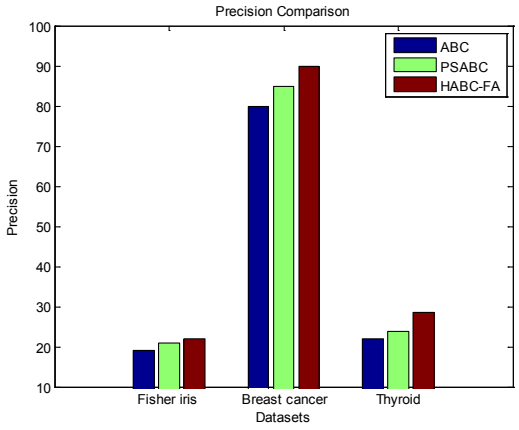

Figure 6. Precision comparison of proposed method Vs. existing method.

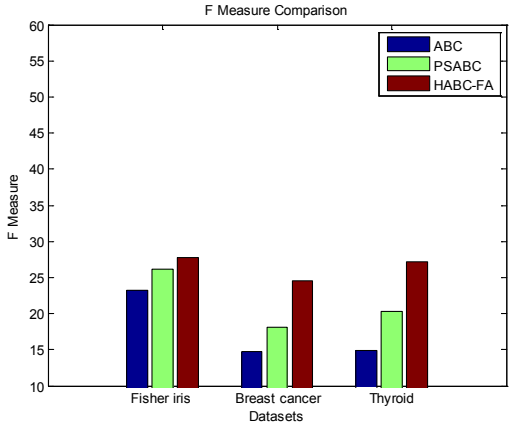

Figure 7. F-measure comparison of proposed method vs. existing method.

6. Conclusion and Future Work

In this paper, a novel clustering algorithm HABC-FA is presented which integrates the fundamental behavior of a Firefly Algorithm combined with ABC, which is to enhance the solution for the complication of clustering. The performance of the newly introduced clustering algorithm is assessed by making a comparison of the performance parameters such recall, precision and f-measures with the existing algorithms using benchmark datasets. The proposed HABC-FA performed well in most of the cases. It can also be observed that the convergence of the HABC-FA is extremely quick for the datasets like iris, cancer and thyroid. The observation from the OLW for VAR metric is that the performance of HABCA-FA is outstanding based on the benchmark datasets. The HABC-FA algorithm has extremely higher probability in discovering the required optimal value in comparison against existing schemes like ABC and PSABC. Therefore, the convergence of the proposed HABC-FA for the Variance measure is quicker as compared to other schemes for most of the benchmark problems. The experimentation results reveal the presented algorithm's performance efficiency. In future work the proposed hybrid clustering method is experiments with high dimensional datasets.

7. References

1. Shelokar PS, Jayaraman VK, Kulkarni BD. An ant colony approach for clustering. Analytica Chimica Acta. 2004; 509(2):187–95.
2. Karaboga K, Dervis D, Akay B. A comparative study of artificial bee colony algorithm. Applied Mathematics and Computation. 2009; 214(1):108–32.
3. Karaboga K, Dervis D, Basturk B. Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems. Foundations of Fuzzy Logic and Soft Computing. Springer Berlin Heidelberg. 2007; 4529:789–98.
4. Kennedy K, James J. Particle swarm optimization. Encyclopedia of Machine Learning. Springer US. 2010; 760–6.
5. Lukasik L, Szymon S, Zak S. Firefly algorithm for continuous constrained optimization tasks. Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. Springer Berlin Heidelberg. 2009; 5796:97–106.
6. Yang Y, Xin-She X. Firefly algorithm, stochastic test functions and design optimisation. International Journal of Bio-Inspired Computation. 2010; 2(2):78–84.
7. Krishnanand KN, Ghose D. Glow worm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. Swarm intelligence. 2009; 3(2):87–124.
8. Das D, Swagatam S. Bacterial foraging optimization algorithm: theoretical foundations, analysis, and applications. Foundations of Computational Intelligence. Springer Berlin Heidelberg. 2009; 3:23–55.
9. Simon S, Dan D. Biogeography-based optimization. IEEE Transactions on Evolutionary Computation. 2008; 12(6):702–13.
10. Yang Y, Xin-She X, Deb S. Engineering optimisation by cuckoo search. International Journal of Mathematical Modelling and Numerical Optimisation. 2010; 1(4):330–343.
11. Tsai T, Wei P. Bat algorithm inspired algorithm for solving numerical optimization problems. Applied Mechanics and Materials. 2012; 148(1):134–7.
12. Yang Y, Xin-She X. Flower pollination algorithm for glob-

al optimization. Unconventional computation and natural computation. Springer Berlin Heidelberg. 2012; 240–9.
13. Karaboga K, Dervis D, Ozturk C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied Soft Computing. 2011; 11(1):652–7.
14. Gao G, Weifeng W, Liu S. Improved artificial bee colony algorithm for global optimization. Information Processing Letters. 2011; 111(17):871–82.
15. Abraham A, Ajith A, Das S, Konar A. Kernel based automatic clustering using modified particle swarm optimization algorithm. In Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, ACM. 2007. p. 2–9.
16. Xu X, Yunfeng Y, Fan P, Yuan L. A simple and efficient artificial bee colony algorithm. Mathematical Problems in Engineering. 2013; 39(3):459–71.
17. Jaeger J, Daniel D. PyG Cluster, a novel hierarchical clustering approach.Bioinformatics. 2014; 30(6): 896–8.
18. Kala K, Rahul K, Shukla A, Tiwari R. A Novel Approach to Clustering using Genetic Algorithm. International Journal of Engineering Research and Industrial Applications. 2010; 3(1):81–8.
19. Nasraoui N, Olfa O, Krishnapuram R. A novel approach to unsupervised robust clustering using genetic niching. Ninth IEEE International Conference on Fuzzy Systems. IEEE. 2000; 1:170–5.
20. Karaboga K, Dervis D, Basturk B. Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. Foundations of Fuzzy Logic and Soft Computing. Springer Berlin Heidelberg. 2007; 789–98.
21. Hyvarinen A, Karhunen J, Oja E. Independent Component Analysis, Wiley Interscience, 2001.
22. Everitt B. Cluster Analysis, Arnold, 3rd edn, 1993.
23. Jong DK. Learning with Genetic Algorithms: An overview. Machine Learning Kluwer Academic publishers. 1988; 3(2):121–38.
24. Vafaie H, Jong KAD. Improving the performance of a Rule Induction System using Genetic Algorithms. Proceedings of the First International Workshop on Multistrategy Learning, Harpers Ferry, W. Virginia, USA. 1991; 1–12.
25. Yu Y, Lei L, Liu H. Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research. 2004; 5:1205–24.
26. Yang XS. Firefly Algorithm, Levy Flights and Global Optimization. Research and Development in Intelligent Systems, Springer Series. 2009; 209–18.
27. Yang XS. Firefly algorithm for multimodal optimization. SAGA. 2009; 5792:169–78.
28. Syed S, Syed A, Senthil Kumaran T. An Energy Efficiency Distributed Routing Algorithm Based on HAC Clustering Method for WSNs. Indian Journal of Science and Technology. 2014 Nov; 7(S7):66–75.
29. Kaliappan V, Thathan M. Enhanced ABC Based PID Controller for Nonlinear Control Systems. Indian Journal of Science and Technology. 2015 Apr; 8(S7):48–56.
30. Vallimeenal R, Rajkumar K. Improving Energy and Network Lifetime using PSO based Apriori in Wireless Sensor Networks. Indian Journal of Science and Technology. 2015; 8(16):1–6.
31. Yazd HGH, Arabshahi SJ, Tavousi M, Alvani A. Optimal Designing of Concrete Gravity Dam using Particle Swarm Optimization Algorithm (PSO). Indian Journal of Science and Technology. 2015 Jun; 8(12):1–10.