Enhanced Artificial bee colony clustering algorithm for mixed data set

C.Nalini*
Department of Information Technology, Kongu Engineering college,Erode
nalinikec@gmail.com

J.Sudeeptha
Associate Software Engineer,Accenture,Chennai
sudeepthajayaraman@gmail.com

Abstract

Data mining techniques are used to extract useful patterns from a large data set.Learning techniques are used to build a good decision making system. K-Means algorithm use Euclidean distance measure to group the data objects. The Euclidean distance measure is sensitive to outliers and is suitable to only numeric values. Real time datasets have missing values and measurements are not in sstandard format. The proposed algorithm expands the capability of K-Means algorithm for clustering mixed dataset employ mixed similarity measure to find the similarity between data objects.Furthermore to improve the performance of clustering analysis, correlation based data imputation is used to impute missing values. Min-Max normalization is used to normalize all values in to a range[0..1].The K-Means algorithm resulted in dead centers, local minima and center redundancy. This problem may arise due to bad initial centers.Artificial Bee Colony Algorithm(ABC) is good in exploration than exploitation.Cooperative search is integrated in ABC to exploit the search space of a bee. In order to enhance data quality and performance of cluster analysis,correlation based data imputation and Min-Max normalization is used to preprocess the data set. In this work elbow method is used to determine number of clusters for the given data set and mixed similarity measure is applied to find similarity between data points. More over integrates the merits of Cooperative search to enhance the search capability of bees. Real time datasets are used to evaluate the outcome of the proposed algorithm.The results demonstrated that the proposed method perform well and generate high quality clusters than other comparative algorithms.

Keywords:*clustering,K-Means,Artificial Bee Algorithm(ABC), cooperative search, Normalization*

1. Introduction

Data mining techniques are used to extract useful insights from a large data set, characterize, discriminate, find relationship and group the objects. Learning algorithms are classified into supervised learning and unsupervised learning algorithm. Learning algorithms are used to group data objects.Clustering is a type of multivariate statistical analysis.The main objective of clustering is to group similar data objects together to assist in understanding the relationships that might exist among them. Clustering algorithms have been applied in various areas like privacy preserving, information retrieval,text analysis, pattern recognition,image processing,video analytics,business intelligence, and medical field etc.,

Clustering algorithms follow two procedures i) hierarchical procedure and ii)non-hierarchical procedure. K-Means is one of the most popular non-hierarchical procedure based algorithm. Bio-inspired optimization algorithms are one of the most popularly used techniques to solve complex problems. The interest in such approaches has been increasing very fast due to their robustness and powerful adaptive search mechanisms and they have been used successfully in many engineering areas for solving difficult multidimensional and multimodal problems. Some of the well-known population based algorithms are Particle Swarm Optimization(PSO), Ant Colony Optimization(ACO), Bacterial foraging strategies, and Artificial Bee colony Optimization algorithm (ABC) etc,. Artificial Bee Colony (ABC) algorithm is developed by Dervis Karaboga motivated by the foraging behavior of honey bees. It is an optimization tool, providing a population-based search procedure in which individuals find new food sources.

2. Related works

Clustering can be used for data exploration,to understand the structure of the data and provides a way to learn the structure of complex data.And also it is useful for detecting outlier in the unlabeled data. The most popular partition clustering(non hierarchical)algorithms are K-Means clustering algorithm and K-Mediods clustering algorithm.These algorithms are not suitable for clustering mixed datasets. These algorithms are sensitive to the cluster centroid initialization and converges to the local optima. Researchers have applied various meta heuristics techniques to solve this problem.The authors [1] reviewed the state-of-the-art in mixed data clustering algorithms, outlined the strength and weakness of each algorithm and recaped the research challenges.And also stated that the necessity of learning algorithm model for mixed data set.Today most of the real world applications produce mixed type of data. K-prototypes clustering algorithm for mixed data sets have proposed in[5]. In this work K-means and K-mediod principles were used to update cluster centers, Euclidean distance used to find similarity between numeric values and used Hamming distance to find similarity between categorical values. W-K-prototypes clustering algorithm[6] extended K-prototypes clustering algorithm.The weights of features are updated based on the importance of the features. The results illustrated that the algorithm performed better than K-prototypes clustering algorithm. K-mean clustering algorithm for mixed data has proposed in [2]. Numeric and categorical features were clustering based on mean and frequency values. The results illustrated that the algorithm performed better than the K-prototypes clustering algorithm. Kernel based clustering algorithm for mixed data proposed in [10]. Hamming distance was used to measure the similarity between categorical attributes and mean values were used to find similarity between numeric attributes.

Some researchers convert categorical features into numerical features before applying K-Means algorithm. In [3], used polar or spherical coordinates for converting categorical values into numeric values.The algorithm produced better results than K-Modes and K-prototypes algorithms. A mutual information based transformation method employed in [11] for the conversion. [13] proposed unsupervised evolutionary clustering algorithm for mixed data. [12] presented Cooperative Artificial Bee Colony (CABC) to solve complex optimization problems. The results showed that CABC performed better than ABC, Particle Swarm Optimization (PSO), and cooperative PSO (CPSO).The feasibility studies exemplified that the importance of an optimized clustering algorithm for mixed data and integration of data cleaning at the initial phase of clustering algorithms.

3. Overview of K-Means clustering and ABC optimization algorithm

3.1 K-Means Clustering algorithm

K-Means clustering algorithm partitioned the data points based on similarity between them.The objective of the algorithm is to increase the similarity between data points within a cluster.

A data set contains “n” data points.K-Means algorithm partition the data points into k clusters, $C_1, \ldots, C_k$, and $C_i \cap C_j = \emptyset$ for $(1 \leq i, j \leq k)$.The algorithm employ Eculidean distance(3.1) measure to find the similarity between a data point and cluster centers.

$$\mathrm{d}(\mathrm{x}_{\mathrm{j}}, c_i) = \left\| c_i - x_j \right\|^2 \qquad (3.1)$$

Where x_j is a data point and c_i is cluster center. The objective of this algorithm is reduce the Squared Error(SE).The quality of a cluster C_i can be measured by using the sum of squared error(3.2) between all objects in C_i and the centroid c_i, defined as

$$\text{Squared\_Error(SE)} = \sum_{i=1}^{k} \sum_{j=1}^{n} \mathrm{d}(\mathrm{x}_{\mathrm{j}}, c_i)^2 \qquad (3.2)$$

The steps of K-Means clustering algorithm are as follows:

Step 1 :Randomly select number of clusters(k) and initialize the cluster centers

Step 2 :Each data point is assigned to the cluster with the nearest center using the Euclidean distance measure

Step 3 :Recalculate the center of each cluster (i.e)
take the mean of data observations which are
grouped in step2.

Step 4 :Repeat step 2 and 3 until there is no change in clusters’ centers or other stopping criteria are met.

3.2 Artificial Bee colony(ABC) algorithm

ABC model has three categories of bees: employed bees, onlookers bees,and scouts bees. The ABC algorithm has four phases (i)intialization phase (ii) Employed bee phase (iii) Onlooker bee phase and (iv) Scout bee phase. The number of employed bees is equal to the number of food sources. Employed bees share their food source information with onlooker bees waiting in the hive. Onlooker bees probabilistically choose their food sources depend on the information shared by employed bees. The abandoned employed bees become scout bee. In ABC model, first randomly select SN solutions(food sources) for creating intial population of size SN*n where SN is equal to number of employed bees and n is the number of dimensions.Each employeed bee X_i (where i=1 to SN) generates a new solution $V_{i.}$.The number of employeed bees and number of onlooker bees is equal to the number of solutions in the search space. The procedure of ABC algorithm as follows:

Step 1: Intialze number employee bees (SN), maximum nuber of iterations, the value of predetermined trials. The initial solution is represented by using the relation(3.3):

$$x_{ij} = x_{j,\min} + (x_{j,\max} - x_{j,\min})*rand \qquad (3.3)$$

Where,

rand is a random number [0…1]

x_{ij} represents the j^{th} dimension of i^{th} solution vector

$x_{j,\min}$ and $x_{j,\min}$ is the minimum and maximum value of j^{th} dimension.

Step 2: Evaluate SN solutions using objective(fitness) function $fit_i(t)$ using (3.4).

$$\mathrm{fit}_i(t) = \begin{cases} \dfrac{1}{1+f_i(t)} & \text{if}(\mathrm{f}_i(t) \geq 0) \\ 1+abs(f_i(t) & \text{otherwise} \end{cases} \qquad (3.4)$$

Step 3: Determine the new solution by iteratively search the solution space.

Step 4: The employed bees update the solutions by using the realtion(3.5)

$$v_{ij} = x_{ij} + r(x_{ij} - x_{kj}) \qquad (3.5)$$

Where,

v_{ij} is a new solution

r is a random number [0…1]

x_{ij} is a old solution

x_{kj} is a randomly selected index.

Step 5: Evaluate the fitness value of new solution and compare with previous fitness value.If it is better than previous value update the fitness value ,otherwise no change in fitness value. The bee memorize the new solution.

Step 6: Now all employed bees share their information with onlooker bees.The onlooker bees probabilitly choose the solution based on their fitness value by using the relation(3.6):

$$p_i = \frac{fitness_i}{summation\ of\ all\ \text{fitness values(SN values)}} \qquad (3.6)$$

Step 7: If there is no updation in the quality of a solution for a predetermined number of iterations then the solution is abandoned. The scout bee randomly choose a new source by using equ and then find a new solution.

Step 8: Repeat the steps from 3-7 until the maximum number of iterations or get convergenced

4. The proposed Cooperative Search based ABC K-Means algorithms for mixed data(CSABC-KM)

CSABC-KM algorithm is a two step process.In order to improve data quality which is fed to the algorithm for building a model, first preprocess the data set and then build a clustering model.

4.1 Data Preprocessing

A real world data may be dirty,incomplete and inconsistent. Data cleaning techniques are used to improve the data quality,accuracy and efficiency of the subsequent mining process.It is used to impute the missing values,smoothing noisy data,identifying outliers, and resolving inconsistencies. Noise is a random error or variance in a measured variable. K-mean clustering algorithm is sensitive to noise and outlier data. This work employs Correlation based data imputation for imputing values,Chi-square test is used for finding correlation between nominal/categorical attributes and Pearson's product moment coefficient for numeric data. Correlation based analysis measure how strongly one attribute implies the other based on the available data. Data transformation strategies are used to standardize the data which is suitable for mining process.There are several methods used to normalize the data such as min-max normalization, z-score normalization, and normalization by decimal scaling.This work applys min-max normalization to normalize the data in the range of [0…1] before clustering the data.

4.2 Clustering Analysis

K-Means algorithm support only numerical attributes.The categorical attributes are converted into numerical attributes.In this case information loss may occurred during conversion. To resolve this problem mixed similarity measrure is used for finding similarity between mixed data obects.ABC algorithm good at exploration and weak in exploitation. it takes more time to obtain converge rate and suffers from premature. Cooperative search is introduced in ABC to achieve good convergence rate.For that the proposed algorithm exploit the globally best(gbest) solution and the best individual solution(lbest) to select best solution for the next iteration.

4.2.1 Mixed similarity measure

In real world application an object is a collection of numerical attributes,categorical attributes, nominal attributes,and ordinal attributes. Finding a similarity between objects of mixed attribute type is difficult.. The modified K-Means algorithm apply mixed distance measure(equ 4.1) to find the similarity between two mixture attribute of data objects. Let the dataset contains p mixed attributes, the similarity between the two objects i and j is defined as:

$$s(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}s_{ij}^{(f)}w_f}{\sum_{f=1}^{p}\delta_{ij}^{(f)}w_f} \qquad (4.1)$$

Where,

The indicator $\delta_{ij}^{(f)}$=0 if x_{if} = x_{jf} =0 and attribute f is asymmetric binary; otherwise $\delta_{ij}^{(f)}$=1.The contribution of attribute f to the similarity between i and j(i.e $s_{ij}^{(f)}$) is computed dependent on its type:

- If f is numeric:
-
$$s_{ij}^{(f)} = \frac{\left| x_{if} - x_{jf} \right|}{\max_h x_{hf} - \min_h x_{hf}} \quad (4.2)$$

Where h runs over all nonmissing objects for attribute f.

- If f is nominal or binary s(i,j)$^{(f)}$ =1 if $x_{if} = x_{jf}$; otherwise 0 The similarity is derived by using the formula (4.3)

$$s(i,j) = \frac{p-m}{p} \quad (4.3)$$

where p is total number of attributes describing an object and m is the number of matches.

- If f is ordinal,compute the ranks r_{if} and z_{if} by using (4.4)

$$z_{if} = \frac{r_{if} - 1}{M_f - 1} \quad (4.4)$$

where M_f is no of states and r_{if} is rank.Optimal weight w_f can be specified in order to raise importace of certain variables that a priori are considered more relevant. If no such preference exit, w_f is set to 1 for all f=1,2,… p and treat z_{if} as numeric.

K-Means algorithm uses iterative relocation technique to improve the partitioning by moving objects from one group to another.It updates the cluster centers by computing the new mean using the objects assigns to the cluster in the previous iteration. The proposed algorithm consider the data type of each attribute to update the cluster center.

4.4. CABC-KM algorithm:

```
Step1 :Data Preprocessing
        Apply correlation based imputation method to impute  missing values
Step 2 : By using Elbow method determine the number of clusters(k).
Step 3: CABC-KM algorithm(k,preprocessed-data set(D))
// Intialization phase:
    ➢ Randomly select k  samples from data set as  cluster center
    ➢ Assume the initial cluster center point as gbest.
    ➢ Initialize iteration=0
    ➢ Send the employed bees to the current food sources
// Employed bees'phase
While (termination condition is not matched)
For each employed bee
      ➢ Compute  similarity measure  by using (4.1)
      End FOR

      ➢ The new solution is found by a bee using (4.4)
```

$$v_{i,d} = x_{i,d} + c_1(x_{i,d} - x_{k,d}) + c_2(x_{i,d} - lb_{j,d}) + c_3((x_{i,d} - gb_d) \quad (4.5)$$

Where,

- $v_{i,d}$,$x_{i,d}$ is new and old position
- lb is the personal best solution

```
        • gb is the global best
        • $c_1, c_2, c_3$ are acceleration coefficients assume as (0.1,0.2,0.2)
    ➢ evaluate fitness value by using(3.2) to find best solution.
    ➢ Check if f(gbest) < f(new_gbest) then Assign gbest=new_gbest.
//Outlooker bees'phase
    ➢ Choose food sources based probability $p_i$ using (3.6)
    ➢ The new solution is selected by bee using (4.4)
    ➢ Check if f(gbest) < f(new_gbest) then assign gbest=new_gbest.
//Scout bees'phase
    ➢ Check the number of food sources in the search space  with threshold value.
    ➢ If there is an employed bee becomes scout then replace it with a new random source
      positions
    ➢ Scout bee select new food sources
    ➢ Memorize the best solution gbest
    ➢ Iteration=iteration+1
End while
Output the best solutions.
```

4 System design and evaluation method

The proposed algorithm is tested on real-life mixed data sets obtained from UCI machine learning repository[4]. The real time datasets namely Horse , Iris,Wine, and Zoo are used to evaluate the performance of proposed method. The performance of the proposed algorithm(CABC-KM) was compared with K-Mean clustering algorithm, PSO based K-Mean clustering algorithm,ABC based K-Mean clustering algorithm.All the algorithms are implemented by using Python. The quality of clustering was analysed by using F-Measure, Average Standard deviation(stdev),Objective function value in best,average and worst values, Davies–Bouldin (DB) index, and Silhouette Coefficient.

- **F-Measure:** It is used to measure the accuracy of clustering algorithms. The F-Measure value indicate the quality of a clustering algorithm.The best clustering algorithm produce highest value.

$$\text{F} - \text{Measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \qquad (5.1)$$

Where

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (5.2)$$

$$\text{recall} = \frac{\text{TP}}{\text{P}} \qquad (5.3)$$

- True Positives (*TP*): refer to the positive tuples that were correctly labeled by the classifier.
- True Negatives(*TN*):refer to the negative tuples that were correctly labeled by the classifier.
- False Positives (*FP*): refer to the negative tuples that were incorrectly labeled as positive.
- False Negatives (*FN*): refer to the positive tuples that were mislabeled as negative.

- **Average standard deviation (stdev):**

$$\text{stdev} = \frac{1}{c}\sqrt{\sum_{i=1}^{n_c} \left\| \sigma(v_i) \right\|} \qquad (5.4)$$

Where,
c is the number of clusters, v_i refers the i^{th} cluster center.

- **Objective function value in best, average and worst values:**

Best denotes the minimum objective function value among all runs, Average indicates the average objective function value of all runs and Worst indicates the maximum value among all times. The higher quality of the clustering algorithm produces smaller value .

- **Davies–Bouldin (DB) index:**

This index aspires the compactness of clusters,use to measure the quality.

$$DBIndex = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\max(md_i + md_j)}{d(c_i, c_j)}\right) \quad \text{where } i \neq j \qquad (5.5)$$

Where k is the number of clusters, md_i is the average distance of members of i^{th} cluster and its center, md_j is the average distance of members of j^{th} cluster and its center and $d(c_i,c_j)$ is the distance between i^{th} and j^{th} cluster centers. The Smaller value indicates a “better”clustering solution.

- **Silhouette Coefficient:**

A data set(D) of n objects and clustered into k partition.For each object x_iεD calculate a (x_i) as the average distance between x_i and all other objects in the cluster to which x_i belongs. Similarly, $b(x_i)$ is the minimum average distance from x_i to all clusters to which x_i does not belong.
The silhouette coefficient of x_i is then defined as

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\left(b(x_i), a(x_i)\right)} \qquad (5.6)$$

The value of the silhouette coefficient is between 0 and 1. The value of $a(x_i)$ reflects the compactness of the cluster to which x_i belongs. The smaller the value specifies compact of the cluster. The value of $b(x_i$) captures the degree to which x_i is separated from other clusters. The larger $b(x_i)$ specifies the separation of x_i from other clusters. The value closer to 1 shows the better performance .

5 Result Analysis

The maximum number of iterations for each algorithm is fixed as 100. Table 6.1 illustrates the simulation results of average of sum of objective function values in best, average and worst values and stdev for 100 runs, Table 6.2 ,6.3, and 6.4 exemplify F-Mesure, Silhouette coefficient,DB index values of algorithms respectively.

The results demonstrated that CABC-KM partition the dataset better than other competitive algorithms. PSO,ABC,CABC-KM algorithms generate the same optimum fitness function value for motor cycle data set.The standard deviation of these algorithms are 13.340,19.118,0.322, the results illustrate that CABC-KM is better than other algorithms.The horse dataset contains missing value and mixed attributes such as numeric,nominal, and categorical.CABC-KM produce the best result in all cases and also the value of stdev is low.It assured that the correlation based imputation method impute appropriate values to missing attributes and also the cooperative search help the bees to exploit the search space in effective way.The zoo dataset containing 17 Boolean-valued attributes and 2 numeric attributes, CABC-KM generate more accurate cluster than other algorithms. Silhouette coefficient and DB index value indicated the compactness of clusters, quality of output, and separation of data objects. The results showed that CABC-KM perform better than other competing algorithms in all data sets. The mixed similarity attribute measure and cooperative search help bees to produce compact and accurate clusters.CABC-KM algorithm generate quality results for all type of data sets.

Table 6.1 Objective function value in best, average and worst values and stdev for 100 runs

Data set	Criteria	K-Means	PSO	ABC	CABC-KM
Motor cycle	Average	3012.300	2060.900	2068.900	2059.700
	Best	2446.300	2060.600	2060.600	**2060.600**
	Worst	4683.200	2110.300	2126.700	2080.400
	Stdev.	439.060	13.340	19.118	**0.322**
Iris	Average	106.05	95.61	94.61	94.60
	Best	97.33	94.60	94.60	96.60
	Worst	120.45	104.93	94.64	94.60
	Stdev.	14.63	1.96	0.01	**0.00**
Wine	Average	18061.00	16302.00	16298.00	16294.00
	Best	16555.0zz0	16292.00	16294.00	**16289.00**
	Worst	18563.00	16384.00	16302.00	16296.00
	Stdev.	793.21	18.27	6.24	**5.47**
CMC	Average	5893.60	5697.40	5695.40	5693.80
	Best	5842.20	5693.80	5693.90	**5693.70**
	Worst	5934.40	5710.70	5698.60	5693.90
	Stdev.	47.17	4.02	1.38	**0.05**
Horse	Average	6591.31	6221.23	6321.12	5876.23
	Best	6532.23	6114.12	6251.12	**5812.56**
	Worst	6612.04	6452.01	6645.78	5902.14
	Stdev.	36.23	10.56	10.68	**1.25**
Zoo	Average	102.36	98.56	98.61	94.6
	Best	96.23	95.25	97.52	**95.06**
	Worst	115.63	105.23	99.63	97.85
	Stdev.	14.23	5.46	1.25	**0.96**

Table 6.2 depicts F-Measure- the accuracy of the model

. Data set	K-Means	PSO	ABC	CABC-KM
Motor cycle	92.25	92.97	93.98	95.65
Iris	95.45	96.32	97.01	98.45
Wine	94.12	94.56	95.21	97.56

CMC	93.21	94.15	94.12	96.23
Horse	92.25	93.23	95.23	98.89
Zoo	93.84	94.25	95.12	97.12

Table 6.3 Illustrates silhouette coefficient

	Silhouette coefficient			
Data set	**K-Means**	**PSO**	**ABC**	**CABC-KM**
Motor cycle	0.4869	0.4841	0.4856	0.5286
Iris	0.5509	0.5366	0.5426	0.6566
Wine	0.5579	0.5548	0.5678	0..6761
CMC	0.4361	0.4386	0.4372	0.5398
Horse	0.3819	0.3912	0.3925	0.5996
Zoo	0.3819	0.3587	0.3871	0.5965

Table 6.4 Illustrates DB index value

Data set	**K-Means**	**PSO**	**ABC**	**CABC-KM**
Motor cycle	5.16	4.69	4.25	4.12
Iris	0.65	0.69	0.51	0.45
Wine	3.49	3.2	3.12	2.92
CMC	8.14	7.85	7.65	5.86
Horse	4.23	4.26	3.25	3.76
Zoo	0.85	1.21	1.32	1.63

6 Conclusion

Clustering analysis partition the data objects based on similarity measure.Missing values affect the performance of clustering algorithms. K-Means algorithm partition the numerical data points into groups.However, the center initialization and fix the number of clusters is a difficult task. To extend the capability of K-Means for mixed data and quality of clustering a cooperative search based ABC –K-Means clustering algorithm has been proposed.The results demonstrated that mixed similarity measure and cooperative search produced accurate and compact clusters for mixed data set.The Horse data set contains missing values.The value of objective function,DB-Index, silhouette coefficient illustrated that CABC-KM algorithm generated more accurate and compact clusters.The results illustrated that the usefulness of correlation based imputation method.Hence CABC-KM algorithm is more suitable for data set which have missing values.

References

[1] Amir Ahmad and Shehroz,(2019),” A Survey of State-of-the-Art Mixed Data Clustering Algorithms”,IEEE Access.

[2] Ahmad and L. Dey, (2007),“A k-mean clustering algorithm for mixed numeric and categorical data,” Data and Knowledge Engineering, vol. 63, no. 2,pp. 503–527.

[3] F. Barcelo-Rico and D. Jose-Luis, (2012)“Geometrical codification for clustering mixed categorical and numerical databases”, Journal of Intelligent Information Systems, vol. 39, no. 1, pp. 167–185.

[4] CL.Blake,CJ.Merz. UCI repository of machine learning databases. Available: http://archive.ics.uci.edu/ml/index.php. University of California, Irvine, Deptartment of Information and Computer Sciences, 1998.

[5] Z. Huang,(1998) “Extensions to the k-means algorithm for clustering large data sets with categorical values,” Data Mining Knowledge Discovery, vol. 2, no. 3, pp.283–304.

[6] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li,(2005), “Automated variable weighting in k-means type clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 657–668.

[7] M. Wei, T. W. S. Chow, and R. H. M. Chan,(2015), “Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation,” Entropy, vol. 17, no. 3, pp. 1535–1548.

[8] Wenping Zou, Yunlong Zhu, Hanning Chen, and Xin Sui1, (2010),” A Clustering Approach Using Cooperative Artificial Bee Colony Algorithm”, Discrete Dynamics in Nature and Society,pp:1-16, doi:10.1155/2010/459796

[9] Z. Zheng, M. Gong, J. Ma, L. Jiao, and Q. Wu,(2010), “Unsupervised evolutionary clustering algorithm for mixed type data,” in IEEE Congress on Evolutionary Computation, pp. 1–8.

[10] Ting Su,Jennifer G. Dy ,(2007),”In search of deterministic methods for initializing K-means and Gaussian mixture clustering”, Intelligent Data Analysis Vol.11, No.4,pp:319-338