Collaborative artificial bee colony k-mean clustering algorithm for mixed data set

C Nalini\textsuperscript{1}, J Sudeeptha \textsuperscript{2}, B S Charulatha \textsuperscript{3} and P Valarmathi \textsuperscript{4}

\textsuperscript{1}Department of Information Technology, Kongu Engineering College, Erode, India nalini_iec@gmail.com \\
\textsuperscript{2}Accenture, Chennai, India \\
\textsuperscript{3}Department of Computer Science and Engineering, Rajalaksmi Engineering College, Chennai, India \\
\textsuperscript{4}Department of Computer Science and Engineering, Mookambigai College of Engineering, Pudukottai, India

Abstract. Data mining techniques are used to extract useful patterns from a large data set. k-mean algorithm is one of the most famous partitioning clustering algorithm. But, Euclidean distance is sensitive to outliers and is suitable to only numeric values. Real time datasets have mixed attribute values, missing values and measurements are not in the standard format. The proposed algorithm extends the ability of the kmean algorithm to use a mixed similarity measure to find the similarity between data objects for clustering mixed datasets. For imputing missing values, correlation based data imputation is used. In addition, k-mean output depends on the initial cluster centre and local optima suffers from the number of clusters(k). In order to improve the efficiency of the k-mean algorithm, Artificial Bee Colony Optimization (ABC) based clustering algorithm is suggested. ABC is successful at exploring the search space, but endures in leveraging the search space. Collaborative search is used to amplify the search quality of bees to amplify the search quality of bees employees. To determine the number of clusters for the given data set, the Elbow method is used. In order to evaluate the outcome of the proposed algorithm, real time datasets are used. The results showed that the proposed method performs well compared to comparative algorithms.

Keywords: Data mining, Clustering, k-mean, Missing values, Correlation imputation, mixed similarity measure, Artificial Bee Algorithm (ABC), Collaborative search, Elbow method

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-mean 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..

2. Related work

For data discovery, clustering may be used to explain the data structure and provide a way to learn the structure of complex data. And it is also helpful in identifying outliers in unlabeled data. K-means clustering algorithms and k-mediods clustering algorithms are the most common partition clustering algorithms. These algorithms are not suitable for clustering mixed datasets. These algorithms are adaptive to the initialization of the cluster centroid and converge to the local optimum.

To solve this problem, researchers have applied different methods of meta-heuristics. [1] analysed the state-of-the-art in mixed data clustering algorithms, outlined the strength and weakness of each algorithm and summarised the research challenges. And also noted that the need for a mixed data set learning algorithm model. Today, most real-world applications generate mixed data form. In[5], the k-prototype clustering algorithm was suggested for mixed data sets. The authors applied Euclidean distance to find similarities between numeric values and used Hamming distance to find similarities between categorical values. k-prototypes clustering algorithm extended by [6] named as W-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. The k-mean clustering algorithm has been proposed by[2] for mixed data. The clustering of numerical and categorical features was focused on mean and frequency values. The results showed that the algorithm worked better than the clustering algorithm for k-prototypes. [7] Proposed kernel-based mixed data clustering algorithm. Hamming distance has been used in this work to calculate the similarity between categorical attributes and mean values used to determine the similarity between numerical attributes.

Researchers transform categorical characteristics into numerical features. In order to transform categorical values into numerical values, Barcelo-Rico et al.(2012) used polar coordinates. The algorithm provided better results than the algorithm for k-modes and K-prototypes. [8] utilized a mutual information based transformation method for data conversion. [10] developed unsupervised evolutionary clustering algorithm for mixed data. [9] presented Cooperative Artificial Bee Colony (CABC) to solve complex problems. The findings showed that CABC performed better than ABC, Particle Swarm Optimization (PSO) and Cooperative PSO (CPSO). The feasibility studies showed that data cleaning techniques and optimised clustering algorithms are important for mixed data.

3. Overview of k-mean clustering and ABC algorithm

3.1 k-mean Clustering algorithm

k-mean 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-mean algorithm partition the data points into k clusters, $C_1$, $C_2$, $C_3$, and $C_i \cap C_j = \emptyset$ for (1 ≤ i, j ≤ k). The algorithm employ Euclidean distance(3.1) measure to find the similarity between a data point and cluster centre.

$$d(x_j, c_i) = \left\| c_i - x_j \right\|^2$$

Where $x_i$ 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
The steps of k-mean clustering algorithm are as follows:

**Step 1**: Randomly select number of clusters (k) and initialize the cluster centres.

**Step 2**: Each data point is assigned to the cluster with the nearest centre 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 step 2.

**Step 4**: Repeat step 2 and 3 until there is no change in clusters’ centre 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) initialization 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 initial population of size SN*n where SN is equal to number of employed bees and n is the number of dimensions. Each employed bee $X_i$ (where i=1 to SN) generates a new solution $V_i$. The number of employed 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**: Initialize number employee bees (SN), maximum number of iterations, the value of predetermined trials. The initial solution is represented by using the relation (3.3):

$$x_{ij} = x_{j,\text{min}} + (x_{j,\text{max}} - x_{j,\text{min}}) \cdot \text{rand} \quad (3.3)$$

Where,
- $\text{rand}$ is a random number $[-1...1]$
- $x_{ij}$ represents the $j^{th}$ dimension of $i^{th}$ solution vector
- $x_{j,\text{min}}$ and $x_{j,\text{max}}$ is the minimum and maximum value of $j^{th}$ dimension.

**Step 2**: Compute objective function and evaluate SN solutions using fitness function $f_i(t)$ using (3.4).

$$f_i(t) = \begin{cases} \frac{1}{1 + f_i(t)} & \text{if } f_i(t) \geq 0 \\ \frac{1 + \text{abs}(f_i(t))}{1 + \text{abs}(f_i(t))} & \text{otherwise} \end{cases}$$

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

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

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

Where,
- $v_{ij}$ is a new solution
- $r$ is a random number $[0...1]$
- $x_{ij}$ is an 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 update the fitness value, otherwise no change in fitness value. The bees memorize the new solution.

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

\[ p_i = \frac{\text{fitness}_i}{\text{sum of all fitness values(SN values)}} \tag{3.6} \]

Step 7: If there is no updating 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 (3.3) and then find a new solution.

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

4. Collaborative Artificial Bee Colony k-mean Clustering Algorithm for Mixed Data Set

Collaborative Artificial Bee Colony k-mean Clustering Algorithm consists of two steps. In order to enhance the data quality that is fed to the model building algorithm, the data set is first pre-processed and then a clustering model is created.

4.1 Data Pre-processing

Data from the real world can be filthy, incomplete and inconsistent. To enhance the data quality, accuracy and effectiveness of the subsequent mining process, data cleaning techniques are used. It is used to impute missing values, smooth noisy data, identify outliers, and resolve inconsistencies. A random error or deviation in a calculated variable is noise. k-mean clustering algorithm is sensitive to noise and outlier data. This study uses the imputation of correlation-based data for imputing values, the Chi-square test is used to identify associations between nominal / categorical attributes and Pearson ‘s product moment coefficient for numeric data. Correlation based analysis is used to calculate how strongly one trait implies the other based on the knowledge available.

4.2 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(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:

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

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:

  \[ s_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\text{max}_h x_{hf} - \text{min}_h x_{hf}} \tag{4.2} \]

  Where h runs over all non missing 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 (4.3)

\[
s(i,j) = \frac{p-m}{p}
\]

(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 \( r_{ij} \) and \( z_{if} \) by using (4.4)

\[
z_{if} = \frac{r_{ij} - 1}{M_f - 1}
\]

(4.4)

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

### 4.3 Find new solution

The new solution is generated by using (4.5)

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

(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)

**Pseudo code of proposed algorithm:**

**Step 1:** Data Pre-processing

Impute missing values using correlation based imputation method

**Step 2:** Determine \( k \)

Find number of clusters \( k \) using Elbow method

**Step 3:**

// Initialization phase

Calculate objective function using (4.1)
Evaluate fitness function (nectar amount) using (3.4)
Select the best value as initial value of gbest
Select top most \( k \) samples as initial cluster centroid
iteration=0
Send the employed bees to the current food sources

// Employed bees’ phase

While (termination condition is not met)

For each employed bee

- Modify the \( j \)th component of the gbest by using the \( j \)th component of bee i
- evaluate fitness value by using(3.2)
- Check if \( f(\text{gbest}) < f(\text{new}_\text{gbest}) \) then
- \( \text{gbest} = \text{new}_\text{gbest} \)
- For employed bee i produce new food source positions \( v_i \) by using (4.5)
- Calculate fitness function
- Apply greedy selection

End for

Calculate probability values for the solutions

//Outlooker bees’ phase

Select food sources based probability \( p_i \) using (3.6)

For each onlooker bee
Modify the $j^{th}$ component of the $\text{gbest}$ by using the $j^{th}$ component of bee $i$

evaluate fitness value by using (3.2)

Check if $f(\text{gbest}) < f(\text{new\_gbest})$ then

$\text{gbest} = \text{new\_gbest}$.

For employed bee $i$ produce new food source positions $v_i$ by using (4.5)

Calculate fitness function

Apply greedy selection

End for

//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

Find new random food source

Memorize the best $\text{gbest}$ solution

Iteration = iteration + 1

End while

Initial solution for k-mean algorithm = \{ $\text{gbest}$ \}
Apply k-mean algorithm

5. System design and evaluation method

Real life mixed data sets obtained from the UCI machine learning repository[4]. Horse, Iris, Wine, and Zoo are the real-time datasets used to test the output of the system proposed. The performance of the proposed algorithm was compared to the k-mean clustering algorithm, the k-mean clustering algorithm based on PSO, and the k-mean clustering algorithm based on ABC. All algorithms are implemented using Python. Using F-Measure, Average Standard Deviation (stdev), Objective Function Value at Best, Average and Worst Values, Davies-Bouldin (DB) Index, and Silhouette Coefficient, the efficiency of clustering was evaluated.

- **F-Measure**: It is used to measure the accuracy of algorithms for clustering. The F-Measure value indicates the quality of a clustering algorithm. The highest value is generated by the best clustering algorithm.

$$F - \text{Measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

(5.1)

Where,

$$\text{precision} = \frac{\text{TP}}{\text{TP + FP}}$$

(5.2)

$$\text{recall} = \frac{\text{TP}}{\text{P}}$$

(5.3)

**True Positive (TP)**: the positive tuples that the classifier has correctly labelled.

**True Negative (TN)**: the negative tuples that the classifier has correctly labelled.

**False Positive (FP)**: the negative tuples that were mislabelled as positive.

**False Negative (FN)**: the positive tuples that were mislabelled as negative.

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

  $$\text{stdev} = \frac{1}{c} \sqrt{\frac{\sum_{c=1}^{c} \sum_{i=1}^{n_c} \| \sigma(v_i) \|^2}{n_c}}$$

(5.4)

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

- **Davies–Bouldin (DB) index:**
This index seeks the compactness of clusters

\[ DBIndex = \frac{1}{k} \sum_{i=1}^{k} \frac{\max(md_i + md_j)}{d(c_i, c_j)} \]  \hspace{1cm} \text{where } i \neq j \quad (5.5) \]

Where k is the number of clusters, md, is the average distance of members of \( i^{th} \) cluster and its center, md, 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 centre. The Smaller value implies a “better” solution for clustering.

- **Silhouette Coefficient:**

A data set \( (D) \) of \( n \) objects and clustered into \( k \) partition. For each object \( x_i \in D \) 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

\[ s(x_i) = \frac{b(x_i) - a(x_i)}{\max(b(x_i), a(x_i))} \]  \hspace{1cm} (5.6) \]

The silhouette coefficient value is between 0 and 1. The \( a(x_i) \) value represents the compactness of the cluster that \( x_i \) belongs to. The smaller the value is defined by the cluster compact. The \( b(x_i) \) value captures the degree by which \( x_i \) is isolated from other clusters. Separation of \( x_i \) from other clusters is defined in the larger \( b(x_i) \). The better result indicates the value closer to 1.

6. Result Analysis

The maximum number of iterations is set as 100 for each algorithm. Table 1 illustrates the simulation results of average of sum of objective function values in best, average and worst values and stdev and Table 2, 3 and 4 indicate F-Measure, silhouette coefficient, algorithm DB index values, respectively.

| Table 1 Objective function value in best, average and worst values and stdev for 100 runs |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Data set**    | **Criteria**    | k-mean          | PSO k-mean      | ABC k-mean      | Collaborative Artificial Bee Colony k-mean |
| 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         |
Table 2 depicts F-Measure- the accuracy of the model

| Data set | K-Means | PSO   | ABC   | Collaborative Artificial Bee Colony k-mean |
|----------|---------|-------|-------|---------------------------------------------|
| 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 3 Illustrates silhouette coefficient

| Data set | Silhouette coefficient | Collaborative Artificial Bee Colony k-mean |
|----------|------------------------|---------------------------------------------|
| 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 4 Illustrates DB index value

| Data set | K-Means | PSO | ABC | Collaborative Artificial Bee Colony k-mean |
|----------|---------|-----|-----|---------------------------------------------|
| 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                                        |

The results showed that the proposed algorithm is better for partitioning the dataset than other competitive algorithms. PSO, ABC, Collaborative Artificial Bee Colony k-mean Mixed Data Set
Clustering Algorithm produces the same optimal motor cycle data set fitness function value. However, the standard deviation of these algorithms is 13,340,19,118,0.322, the results show that the proposed algorithm is better than other algorithms. The horse dataset includes missing value and mixed attributes such as numeric, nominal, and categorical. In all situations, the proposed algorithm provides the best result and the value of stdev is also low. It ensures that the correlation-based imputation approach imputes sufficient values to missing attributes and also allows the bees to efficiently leverage the search space. The proposed algorithm produces more precise clusters than other algorithms, with a zoo dataset containing 17 Boolean-value attributes and 2 numeric attributes. The silhouette coefficient and the DB index value showed cluster compactness, performance efficiency, and data object separation. The results showed that the Collaborative Artificial Bee Colony k-mean Mixed Data Set Clustering Algorithm performs in all data sets better than other competing algorithms and achieves consistency results for all types of data sets.

7. Conclusion
The correlation-based data imputation approach was used to impute missing values to enhance the data quality. To extend the capability of the k-means clustering algorithm to mixed data, a mixed similarity measure has been proposed. To improve the effectiveness of the k-mean clustering algorithm, the Collaborative Artificial Bee Colony based k-mean clustering was developed. The results showed that the collaborative search enhanced the search capacity of the bees and delivered precise and compact clusters.

References
[1] Amir Ahmad and Shehroz,(2019),” A Survey of State-of-the-Art Mixed Data Clustering Algorithms”, IEEE Access. vol. 7, pp. 31883-31902, doi:10.1109/ACCESS.2019.2903568.
[2] Ahmad and Lipika 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] Barcelo-Rico F and Diez Jose-Luis, (2012)“Geometrical codification for clustering mixed categorical and numerical databases”, Journal of Intelligent Information Systems, vol. 39, no. 1, pp. 167–185 https://doi.org/10.1007/s10844-011-0187-y
[4] Blake CL,Merz CJ, 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] Huang J Z,(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] Huang J,Z,Michael .K. Ng, Rong H, and Li Z,(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] 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
[8] Wei M, Chow T. W. S., and Chan R. H. M.,(2015), “Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation,” Entropy, vol. 17, no. 3, pp. 1535–1548.
[9] 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
[10] Zheng Z, Gong M, Ma J, Jiao L, and Wu Q,(2010), “Unsupervised evolutionary clustering algorithm for mixed type data,” in IEEE Congress on Evolutionary Computation, pp. 1–8.