Multi-objective C-means Data Clustering Algorithm using Self-Adaptive Differential Evolution

Omar S. Soliman , Doaa A. Saleh

Abstract— : This paper proposes a Multi-objective C-means Data Clustering algorithm using Self-Adaptive Differential Evolution (DE) for improving the performance of data clustering by introducing three data clustering validity indices.. The proposed algorithm composed of three objectives: including the symmetry-index to maximize similarity within clusters, the compactness index to maximize dissimilarity among clusters, and validity Silhouette index to improve the validity of data clustering. Self- adaptive DE is similar to the traditional DE algorithm except two changes in the mutation and the crossover operations [19], where DE is a global optimization technique [13]. The proposed algorithm is implemented and evaluated using twenty benchmark data sets and compared with different 5 data clustering algorithms that MOSAC-Means, GenClustMOO, MOCK, VGAPS, and GenClustPESA2. The experimental results showed that the proposed algorithm is performing well compared with the previous algorithms.

Index Terms— Multi-objective Data Clustering, Self-Adaptive Differential Evolution, Symmetry Index, Compactness Index, Silhouette Index.

1. Introduction

Validity clustering indexes are important to evaluate the performance of the tested data clustering algorithm [3]. Recently, some studies proposed the validity indexes as objective functions in a multi-objective framework [23]. Most of these studies were based on K-means for numerical data. Multi-objective clustering aims to improve the performance of data clustering through achieving some conflicted objective functions. Most time, these conflicted functions are from validity clustering indexes. In multi-objective clustering, the goodness of each cluster should be judged not only by the clustering algorithm that generated it, but also by external and/or internal assessment criteria [20], [38]. Validity indexes [6], [14], [15], [16], [18], [22], [39] are divided into two main categories that internal and external validity indexes based on internal criteria and external criteria. Bic-index, Calinski-Harabasz index, Davies-Bouldin index, Silhouette index, Dunn index, and NIVA index are considered as internal indexes. But, F-measure, Purity, Precision, Recall, Minkowski score, and Adjust Rand Index are also examples of external validity indexes. Evolutionary algorithms [7], [12] are considered very successful in carrying out multiple objectives optimization (MOO). In addition, most evolutionary algorithms are robust and multi-modal which proves to be a distinct advantage in the solution of such problems. Optimization of multiple objectives requires that the relative importance of each objective be specified in advance which requires a prior knowledge of the possible solutions.

MOO is used when dealing with the real-world problems where there are several objectives that should be optimized simultaneously.
In general, a MOO algorithm usually admits a set of solutions that are not dominated by any solution. During recent years, many multi-objective evolution algorithms, such as multi-objective EA (MOEA), have been suggested to solve the MOO problems.

Differential evolution (DE) is a branch of evolutionary algorithms developed for optimization problems over continuous domains. In DE, each variable is represented in the chromosome by a real number. Furthermore, DE is also considered one of the class of genetic algorithms (GAs) which use the same operations of crossover, mutation, and selection on a population in order to minimize an objective function over the course of successive generations [5], [13]. In [19], Self- adaptive DE algorithm is the same as the traditional DE algorithm except two changes in the mutation and the crossover operations**.**

This paper introduces a new multi-objective data clustering algorithm for improving the performance of data clustering based on Self-Adaptive-DE algorithm (MODEC-Means). The reset of the paper is organized as follows: section 2 &3 introduce related works of multi-objective data clustering and background for Self-Adaptive DE. Section 4 presents the Multi-objective mathematical model; Section 5 introduces the proposed

- *Omar S, Soliman is currently an Associate Professor, Faculty of Computers and Information, Cairo University, Egypt.*
 E-mail: Dr.omar.soliman@gmail.com

- *Doaa A. Saleh is currently PhD* Candidate *, Faculty of Computers and Information, Cairo University, Egypt.*
. E-mail: d.saleh@fci-cu.edu.eg

algorithm; Section 6 presents used data sets, experimental results, discussion and analysis of obtained results; where the last section is devoted to the conclusion.

2. BACKGROUND AND RELATED WORK

Clustering is an important real world problem and several clustering algorithms usually attempt to optimize some validity measure such as the compactness of the clusters, separation among the clusters or combination of both. Therefore, it is better to optimize compactness and separation separately rather than combining them in a single measure to be optimized. In [1], authors introduced a multi-objectives categorical data clustering model around medoids by using two objective functions. These two objectives are that K-medoids error function and Silhouette validity index which have been simultaneously optimized using multi-objective GA. In [9], multi-objective DE crisp clustering algorithm was developed for categorical using K- medoids.

Several measures are proposed to evaluate the performance of data clustering algorithms. Therefore, objective functions are different in each study to handle data clustering under multi-objective framework. In [8], they evolved DE data clustering algorithm with two objective that the Xie-Beni index and Euclidean distances. In [37], authors selected two complementary objectives that compactness and connectedness of clusters based on AIS. In [27], symmetry-index and Euclidean distances are used as objective functions and solved by SA. In [10], a new Dynamic Multi-objective Differential Crisp Clustering algorithm was proposed. That algorithm has two conflicting objective functions that DB index and CS measure for finding global compactness and separation among the clusters.

In [28], they proposed a Multi-objective C-means data clustering selected using used SA, these objectives was symmetry-index, connectively-index, and I-index In [13], authors selected the Xie-Beni index XBq and a penalized version as the two objectives based on DE and the FCM function Jq. In [26] and [25], symmetry- index and average of symmetry- index have been used as objective functions for achieving stability among clusters. In [2], authors selected two objectives that the Xie-Beni index and Euclidean distances based on FPSO. Multi-objectives data clustering algorithm was proposed in [24] with four objective functions including total compactness of the partitioning, total symmetry present in the clusters, cluster connectedness, and Adjust Rand Index using Hybrid Intelligent Systems (HIS).

3. SELF-ADAPTIVE DIFFERENTIAL EVOLUTION

DE is a population-based global optimization algorithm that uses a real-coded representation [5]. DE is also considered one of the class of genetic algorithms (GAs) which use the same operations of crossover, mutation, and selection on a population in order to minimize an objective function over the course of successive generations [13]. Self- adaptive DE algorithm is the same as the traditional DE algorithm except two changes in the mutation and the crossover operations [19]:

1) In the mutation, the step length (F) will be adapted based on a cauchy distribution with fixed mean μ and adaptive scale parameter δ as follows:

$$F_{i,t+1} = \begin{cases} c(\mu,\delta_{i,t+1}), & if\ rand_1 \le \pi_1 \\ c(\mu,\delta_{i,t}) & otherwise \end{cases} \qquad (1)$$

where: $\delta_{i,t+1} = \delta_l + \delta_u * rand_2$

2) In the crossover, the change is in calculating the control parameter CR instead of being a constant, where

$$CR_{i,t+1} = \begin{cases} rand_3 & if\ \ rand_4 \le \pi_2, \\ CR_{i,t} & otherwise \end{cases} \qquad (2)$$

Where δ_l *and* δ_u are the lower and upper bounds to the scale parameter δ respectively, *rand*$_j$ ∊ [0, 1], j = 1, 2, 3, 4 are uniform random numbers, and π_1 and π_2 represent the absolute probabilities to adapt *F* and *CR* respectively.

4. MATHEMATICAL MODEL

The proposed algorithm has three objective functions in addition to three constrain as shown in figure 1. These three objective functions reflect three different aspects of good clustering solutions. The first one quantifies the amount of symmetry present in a particular partitioning, the second one minimizes the connectedness among data clusters, and the third one measures the goodness/ or validity of overall data clustering performance.

$$Maximize \quad MOP = (sym(K), \frac{1}{Con(K)}, Sil(K)) \qquad (3)$$

subject to

$$\sum_{j=1}^{n} \mu_{kj} \ge 1 \quad for\ \ k = 1, ..., k \qquad (4)$$

$$\sum_{k=1}^{K} \mu_{kj} = 1 \quad for\ \ j = 1, ..., n \qquad (5)$$

$$\sum_{k=1}^{K} \sum_{j=1}^{n} \mu_{kj} = n \qquad (6)$$

Figure 1 The proposed multi-objective data clustering algorithm

In eq. (3), *Sym*, *Con*, and *Sil* refer to symmetry-index, compactness index, and validity Silhouette index, respectively. Given the system constraints as *eq. 4, 5, and 6,*
Where *n* is a number of data points; *k* refers to number of data clusters, and $\mu_{kj} \in [0, 1]$ is the membership of pattern x_i to cluster C_k .In crisp lustering*:*

$\mu_{kj} = 1 \; if \;\; x_i \in C_k$, otherwise μ_{kj} =0.

The objective functions in the above *MOP* are described as follows:

1) The first objective function is a cluster validity Index (*Sym-index)* that identifies the total compactness of the partitioning based on the Euclidean distance [33]. It is defined as: Let $\bar{x}$ be a point, the reflected symmetrical the point of $\bar{x}$ with respect to a particular center $\bar{c}$ is $2 \times \bar{c} - \bar{x}$. Let us denote this by $\bar{x}^*$. Let the first and the second unique nearest neighbors to $\bar{x}^*$ be at Euclidean distances of d_1 and d_2, respectively. Then $d_{ps}(\bar{x}, \bar{c}) = \frac{d_1 + d_2}{2} \times d_e(\bar{x}, \bar{c})$, where $d_e = (\bar{x}, \bar{c})$ is the Euclidean distance between the point $\bar{x}$ and $\bar{c}$

$$Maximize \quad Sym(k) = (\frac{1}{k} \times \frac{1}{\varepsilon_k} \times D_k) \qquad (7)$$

$where \; \varepsilon_k = \sum_{i=1}^{K} E_i, \; and \; E_i = \sum_{j=1}^{n_1} d^*_{ps}(\, x^i_j, \bar{c_i})$, and D_k is the maximum matching distance between two cluster centers among all pairs of centers.

2) The second objective function is a connectivity based cluster validity index (*Con-index)* that reflecs the total symmetry of the clusters (connectively-index) [38]. This objective is defined as: let $\overline{m_k}$ is the medoid of the kth cluster, it is the point of that cluster which has the minimum average distance to all the other points in that cluster.

$\overline{m_k} = \arg\min_{i=1}^{n_k} \frac{\sum_{j=1}^{n_k} d_e(\overline{x_i^k}, \overline{x_j^k})}{n_k}$, where $d_e(\overline{x_i^k}, \overline{x_j^k})$ is calculated by the Euclidean distance, n$_k$ is the total number of data points in the cluster kth , $\overline{x_i^k}$ refers to the data point ith in the cluster kth, then $\overline{m_k} = \overline{x^k_{\min index}}$. Then the Conn-Mod-index function will be as follows:

$$Minimize \quad Con = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_k} d_{short}(\overline{m_i}, \overline{x_j^i})}{n \times \min_{i,j=1 \wedge i \neq j}^{k} d_{short}(\overline{m_i}, \overline{m_j})} \qquad (8)$$

where d$_{short}(\overline{m_i}, \overline{x_j^i})$ and d$_{short}(\overline{m_i}, \overline{m_j})$ will be calculated by the Euclidean distance. The smaller values of Con-index correspond to good partitioning. Furthermore, achieving the good partitioning, the value of Con-index has to be minimized.

3) The third objective function is the Silhouette Validity Index for computing the silhouette width for each data point, average silhouette width for each cluster and overall average silhouette width for the total data set [39]. Silhouette Validity Index is considered from the internal validity indexes. The silhouettes width of ith data point is computed by following this formula

$$S_i = \frac{b_i - a_i}{\max\{b_i - a_i\}}, \quad -1 \le q_i \le 1$$

where *a(i)* is average dissimilarity of ith data point to all other points in the same cluster; and *b(i)* is minimum of average dissimilarity of ith data point to all data points in other cluster. A value of S_i is between -1 and 1 when it close to 1 indicates that the data point is assigned to a very appropriate cluster. When S_i is close to zero, it means that data point could be assign to another closest cluster as well because it is equidistant from both the clusters. But, if *Si* is close to –1, it means that data is misclassified and lies somewhere in between the clusters. The overall average silhouette width for the entire data set is the average *Si* for all data points in the whole dataset. The largest overall average silhouette indicates the best clustering. Therefore, the number of cluster with maximum overall average silhouette width is taken as the optimal number of the clusters.

$$Maximize \quad Sil(k) = \frac{1}{N} \sum_{i=1}^{N} S_i \qquad (9)$$

5. Proposed Algorithm

The proposed MODEC-Means algorithm is developed for hybrid Multi-objective data clustering based on Self-Adaptive DE. This algorithm is divided into several steps as follow; the steps (1 and 2) are an initialization steps for some required parameters; rest of the steps are considered the main body of the proposed algorithm which start with calculating the centroids matrix for each individual then distance and membership matrices should be updated, the next two steps are applying the adapted mutation, then crossover operators, and the next step evaluate the candidate C for each parent. After that, the selection operator should be applying to create the new population based on fitness function. The proposed algorithm is described in figure 2.

```
1. Initialize the population P=100of random individuals.
2. Initialize the parameters F=2, CR=0.7, p=2
3. While stopping criterion not met, do:
  3.1. For each individual Pi (i = 1. . . NP) from P repeat:
      a) Calculate the centroid for each individual in the population.
      b) Update distance between data objects and the new centers of clusters by Euclidean
         distance (ED).
      c) Calculate distance among the different clusters by ED.
      d) Apply adapted mutation operator eq. (1) →DE/rand/1
      e) Apply adapted crossover operator eq. (2)
      f) Evaluate fitness of the candidate C from parent Pi for each objective function.
```

```
        g) Apply selection operator to create new-population by comparing each candidate
           C with its parent P according to: If the candidate dominates the parent, the
           candidate replaces the parent. If the parent dominates the candidate, the
           candidate is discarded. Otherwise, the candidate is added in the population.
    3.2. If the population has more than pop Size individuals, truncate it
    3.3. Randomly enumerate the individuals in P.
    3.4. If not met stopping criterion, go to step (3.1)

4. Determine a set of non-dominate solutions (individuals) from new-population.
5. Select optimal solution from a set of non-dominate solutions according to ARI measure
   performance.
```

Figure 2 MODEC-means Algorithm

Selection operator compares between candidate and its parent based on domination for each objective function in them. For purpose of maximization, if the fitness values of the three objective functions in candidate C are greater than the values of all three objective functions in parent P, then candidate C dominates parent P, and vice true. Finally, we have to determine a set of non-dominate solutions (individuals) from new-population, then select optimal solution from a set of non-dominate solutions according to Adjust Rand Index (ARI) [21], [22] measure performance. Next section will discussed used data sets, obtained results, and analysis under an umbrella of experimental results.

6. EXPERIMENTAL RESULTS

The proposed algorithm is compared with five different algorithms MOSAC-Means GenClustMOO, MOCK, VGAPS, and GenClustPESA2 based on F-Measure. GenClustMOO [28] was developed in 2013 based on Simulated Annulling (SA) and achieved good results. Therefore, we need firstly to hybrid SA [34], [35] with the mathematical model (in fig. 1) for ensuring the quality of the proposed model with the same values of parameters which used in ref. [28]. Secondly, the proposed MODEC-Means algorithm is developed based on Self-Adaptive DE to get better results, where DE is a global optimization technique.

6.1. Used Data Sets

The proposed algorithm is implemented on twenty benchmark data sets by using VC++ and evaluated by a performance measure (F-Measure). Used data sets are divided into four groups, as follows:

- Group_1: consists of four data sets with symmetrical shaped clusters that: (Sym3-2& Ellip2-2) used in [17], and (ring3-2& rect3-2) found in [16].
- Group_2: contains of five data sets with hyper-spherical shaped clusters found in [32], [36], and [29] that: (Sph 5 2, Sph 4 3, Sph 6 2, Sph 9 2, and Sph 10 2).
- Group_3: consists of six data sets with well-separated clusters that: Pat1 used in [30], Pat2 used in [31], and (Size5, Square4, Twenty, Forty) found in [11].
- Group_4: are five real life datasets [4] that (Iris, Wine, Liver-Diseases, Lung-Cancer, and Glass).

6.2. Experimental Results

The obtained results of the 30 independent runs are summarized and tabulated in tables 1 and 2. Table 1 contains the best result in the thirty runs and the computed rank (the numbers in between brackets). The proposed algorithm is compared with five different algorithms that MOSAC-Means, GenClustMOO, MOCK, VGAPS, and GenClustPESA2 algorithms based on a performance measure (F-measure). Firstly, the MOSAC-Means is performed well with comparing of GenClustMOO algorithm according to F-measure.

The rank is also computed according to this performance measure. This rank is taking values from 1 to 6, where the best will get rank with value one and the worst will take five. The optimal number of clusters is represented in figure 3 on different 20 data sets for the proposed algorithm and other compared data clustering algorithms. In table 1, the proposed algorithm mostly improved in the values of F-measure except in Iris, Wine, Lung Dis. Data sets. The MODE_C-means algorithm gets 1.15 in the rank average, then comes the MOSAC-Means algorithm and GenClustMOO with a few differences to be 2.1 and 2.3 with respectively, then GenClustPESA2 has 2.85, and finally MOCK and VGAPS have the value in the rank average to be 3.3 and 3.35 with respectively. In table 2, the proposed algorithm mostly progressed in F-measure values except in some data sets Ring, Ellips, Sph10-2, Size-5, Iris, Wine, and Lung with a few difference.

Table 1: The best results of F-measure & computed rank through the thirty runs for 20 different data sets by the proposed algorithm and other five algorithms.

Data sets	*MODE C-Means*	*MOSAC -Means*	*GenClustMOO*	*MOCK*	*VGAPS*	*GenClustPESA2*
Sym _5-2	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)
Ellip_2-2	1.00(1)	1.00(1)	0.971(3)	0.667(5)	1.00(1)	0.968(4)
Ring_3-2	0.966(1)	0.958(4)	0.964(2)	0.801(5)	0.961(3)	0.961(3)
Rect_3-2	1.00(1)	1.00(1)	1.00(1)	1.00(1)	0.736(2)	1.00(1)
Sph_5-2	0.978(1)	0.943(2)	0.957(1)	0.902(4)	0.541(5)	0.936(3)
Sph_4-3	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)
Sph_6-2	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)	1.00(1)
Sph_10-2	0.983(1)	0.969(3)	0.981(2)	0.717(6)	0.752(5)	0.931(4)
Sph_9-2	0.791(1)	0.783(2)	0.681(4)	0.717(3)	0.481(6)	0.652(5)
Pat1	0.967(1)	0.952(2)	0.946(3)	0.547(4)	0.418(5)	0.946(3)
Pat2	1.00(1)	1.00(1)	1.00(1)	0.545(3)	0.582(2)	1.00(1)
Size5	0.984(1)	0.966(3)	0.968(2)	0.791(6)	0.816(5)	0.883(4)
Square4	0.983(1)	0.923(2)	0.918(4)	0.895(5)	0.925(3)	0.878(6)
Twenty	1.00(1)	1.00(1)	1.00(1)	1.00(1)	0.479(3)	0.948(2)
Forty	1.00(1)	1.00(1)	1.00(1)	1.00(1)	0.095(3)	0.979(2)
Iris	0.803(2)	0.761(4)	0.788(3)	0.775(5)	0.754(6)	0.926(1)
Wine	0.724(2)	*0.711*(3)	0.709(4)	0.726(1)	0.617(5)	0.437(6)
Liver Dis.	0.718(1)	*0.720*(2)	0.673(4)	0.671(5)	0.705(3)	0.603(6)
Lung Can.	0.817(2)	*0.799*(4)	0.802(3)	0.443(6)	0.741(5)	0.843(1)
Glass	0.542(1)	*0.501*(3)	0.494(4)	0.534(2)	0.534(2)	0.534(2)
Rank Average	1.15	2.1	2.3	3.3	3.35	2.85

Table 2: The average value of the F-measure and standard deviation through the thirty runs on the different 20 data sets for the proposed algorithm and five different data clustering algorithms

Data sets	***MODE C_means***	***MOSA C-Means***	***GenClustMOO***	***MOCK***	***VGAPS***	***GenClustPESA2***
Sym _5-2	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
Ellip_2-2	0.988 ± 0.021	0.974 ± 0.012	0.971 ± 0.01	0.667 ± 0.014	**1.00** ± 0.0101	0.968 ± 0.001
Ring_3-2	0.961 ± 0.01	0.948 ± 0.027	**0.964** ± 0.021	0.801 ± 0.011	0.961 ± 0.013	0.961 ± 0.021
Rect_3-2	**1.00 ± 0.01**	**1.00 ± 0.01**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.736 ± 0.011	**1.00 ± 0.00**

Sph_5-2	**0.972 ± 0.01**	0.9402 ± 0.017	**0.957** ± 0.021	0.902 ± 0.0113	0.541 ± 0.011	0.936 ± 0.012
Sph_4-3	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
Sph_6-2	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
Sph_10-2	0.971 ± 0.03	0.967 ± 0.011	**0.981** ± 0.011	0.717 ± 0.013	0.752 ± 0.0109	0.931 ± 0.021
Sph_9-2	**0.749** ± 0.11	**0.688** ± 0.21	0.681 ± 0.012	0.717 ± 0.009	0.481 ± 0.012	0.652 ± 0.018
Pat1	**0.951** ± 0.11	0.949 ± 0.013	0.946 ± 0.013	0.547 ± 0.011	0.418 ± 0.014	0.946 ± 0.009
Pat2	**1.00** ± 0.029	**1.00** ± 0.029	**1.00** ± 0.012	0.545 ± 0.013	0.582 ± 0.021	**1.00** ± 0.00
Size5	0.963± 0.01	0.963± 0.023	**0.968** ± 0.001	0.791 ± 0.012	0.816 ± 0.013	0.883 ± 0.011
Square4	**0.978** ± 0.21	0.9173 ± 0.017	0.918 ± 0.014	0.895 ± 0.011	0.925 ± 0.013	0.878 ± 0.011
Twenty	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.479 ± 0.022	0.948 ± 0.015
Forty	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.95 ± 0.006	0.979 ± 0.015
Iris	0.779 ± 0.011	0.759 ± 0.021	0.788 ± 0.011	0.775 ± 0.022	0.754 ± 0.013	**0.926** ± 0.015
Wine	0.718 ± 0.016	0.708 ± 0.07	0.709 ± 0.012	**0.726** ± 0.002	0.617 ± 0.008	0.437 ± 0.012
Liver Dis.	**0.714** ± 0.07	0.7011 ± 0.09	0.673 ± 0.002	0.671 ± 0.012	0.705 ± 0.009	0.603 ± 0.015
Lung Can.	0.7936 ± 0.017	0.766 ± 0.021	0.802 ± 0.014	0.443 ± 0.011	0.741 ± 0.008	**0.843** ± 0.002
Glass	**0.523 ± 0.01**	0.498 ± 0.011	0.494 ± 0.012	0.534 ± 0.006	0.534 ± 0.008	0.534 ± 0.012

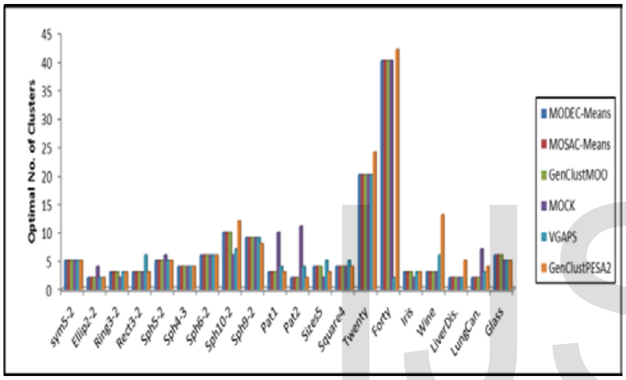

Figure 3 The optimal no. of clusters on all data sets for the proposed algorithm and other compared data clustering algorithms

Figure 3 shows the optimal numbers of clusters for each compared algorithm through the 20 data sets. In the most data sets, the optimal numbers of clusters are convergent except in *sph10-2, pat1, pat2, twenty, forty, wine, and lung cancer* data sets. From the experimental results, MOSAC-Means and MODEC-Means algorithms achieved the better results comparing with the other algorithms. Therefore, figures 4 and 5 display the behavior of the proposed algorithms and MOSAC-Means through the average and the best values of 30 independent runs. We can observe from these figures that MODEC-Means improved in the performance of data clustering for the used data test.

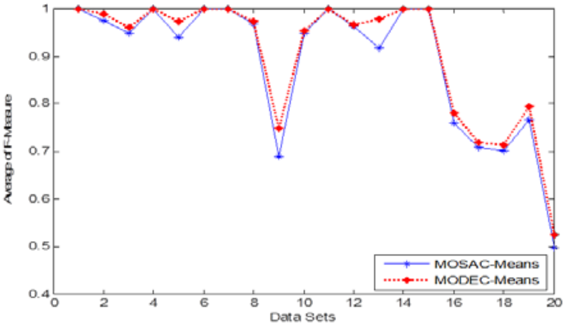

Figure 4 The average value of 30 runs of F-measure for Algorithms (MOSAC-Means) *vs.* (MODEC-Means) through 20 data sets

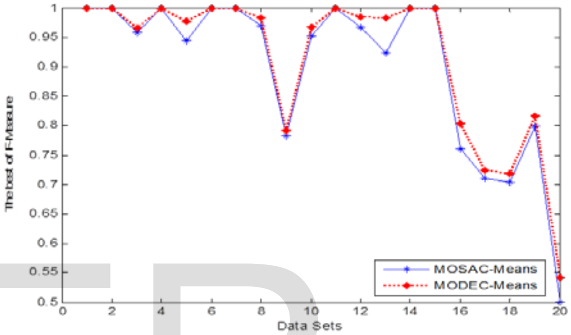

Figure 5 The best value of 30 runs of F-measure for Algorithms (MOSAC-Means) *vs.* (MODEC-Means) through 20 data sets

7. CONCLUSION

This paper introduced an algorithm for improving the performance of data clustering through introducing three data clustering validity indices. These clustering validity indices is modeled as a multi-objective data clustering based on Self-Adaptive DE. The three objectives are a symmetry-index to maximize similarity within clusters, the compactness index to maximize dissimilarity among clusters, and the validity Silhouette index to improve the validity of data clustering. The proposed algorithm was implemented on twenty benchmark data sets and compared with five different data clustering algorithms MOSAC-Means, GenClustMOO, MOCK, VGAPS, and GenClustPESA2. The obtained results showed that the proposed MODEC_Means algorithm performed well compared with its compared algorithms.

References

[1] A. Mukhopadhyay, U. Maulik, Multiobjective Approach to Categorical Data Clustering, IEEE Congress on Evolutionary Computation, 2007.

[2] B. A. Attea, A fuzzy multi-objective particle swarm optimization for effective data clustering, Memetic Comp. (2010) 2:305–312.

[3] B. B. Baridam, More Work on *K* -Means Clustering Algorithm: The Dimensionality Problem, *International Journal of Computer Applications (0975 – 8887) Volume 44– No.2, April 2012*.

[4] Blake, C., & Merz, C. (1998). UCI Repository Machine Learning Datasets.

[5] D. Ardia, K. Boudt, P. Carl, K. M. Mullen, B. G. Peterson, Differential Evolution with DE optima, The R Journal Vol. 3/1, June 2011.

[6] E. Rendón, I. Abundez, A. Arizmendi, E. M. Quiroz, Internal versus External cluster validation indexes, International Journal of Computers and Communications, Issue 1, Volume 5, 2011.

[7] I. Saha, D. Plewczy´nski, U. Maulik, S. Bandyopadhyay, Consensus Multiobjective Differential Crisp Clustering for Categorical Data Analysis, RSCTC 2010, LNAI 6086, pp. 30–39, 2010.

[8] I. Saha, U. Maulik, D. Plewczy´nski, Multiobjective Differential Crisp Clustering for Evaluation of Clusters Dynamically, Springer-Verlag Berlin Heidelberg 2011, Man-Machine Interactions 2, AISC 103, pp. 307–313.

[9] I. Saha, A. Mukhopadhyay, Improved Crisp and Fuzzy Clustering Techniques for Categorical Data, IAENG International Journal of Computer Science, 2008, 35:4, IJCS_34_4_01.

[10] I. Saha, U. Maulik, D. Plewczynskia, A new multi-objective technique for differential fuzzy clustering, Applied Soft Computing 11 (2011) 2765–2776.

[11] J. Handl, J. Knowles, Evolutionary multiobjective clustering. (PPSN VIII). Pages 1081-1091. LNCS 3242, 2004.

[12] J. Handl, J. Knowles, (2007) An evolutionary approach to multiobjective clustering. IEEE Transactions on Evolutionary Computation 11(1):56-7

[13] K. Suresh, D. Kundu, S. Ghosh, S. Das, A. Abraham, S. Y. Han, Multi-Objective Differential Evolution for Automatic Clustering with Application to Micro-Array Data Analysis, ISSN 1424-8220, Sensors 2009.

[14] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. R. Dougherty, Model-based evaluation of clustering validation measures. Pattern Recognition, 40(3):807–824, 2007.

[15] M. Rawashdeh, A. Ralescu, Crisp and Fuzzy Cluster Validity - Generalized Intra-Inter Silhouette Index, 978-1-4673-2338, 2012 IEEE.

[16] M. C. Su, C. H. Chou, and C. C. Hsieh, “Fuzzy C-Means Algorithm with a Point Symmetry Distance,” International Journal of Fuzzy Systems, vol. 7, no. 4, pp. 175-181, 2005.

[17] M. C. Su and C. H. Chou, “A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 674-680, June 2001.

[18] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. rez, I. Perona, An extensive comparative study of cluster validity indices, Pattern Recognition 46 (2013) 243–256.

[19] O. S. Soliman, L. T. Bui, A self- Adaptive Strategy for Controlling Parameters in Differential Evolution, IEEE, Congress on Evolutionary Computation, pp. 2837-2842, 2008

[20] Q. Zhao, M. Xu, P. Fränti, Sum-of-Squares Based Cluster Validity Index and Significance Analysis, ICANNGA 2009, LNCS 5495, pp. 313–322, 2009.

[21] Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods", Journal of the American Statistical Association, **66**, 846–850.

[22] R.J.G.B. Campello, A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment, Pattern Recognition Letters 28 (2007) 833–841.

[23] S. Saha, A. Ekbal, K. Gupta, S. Bandyopadhyay, Gene expression data clustering using a multi-objective symmetry based clustering technique, Computers in Biology and Medicine, Vol. 43(11), pp. 1965-1977, 2013.

[24] S. Saha, A. Ekbal, A. K. Alok, Semi-supervised clustering using multi-objective optimization, Hybrid Intelligent Systems (HIS), the 12th International Conference on IEEE, (2012), pp. 360 -365

[25] S. Bandyopadhyay, Multi-objective Simulated Annealing for Fuzzy Clustering With Stability and Validity, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE 2011, Vol. 41(5), pp. 682 – 691.

[26] S. Saha, S. Bandyopadhyay , A new multi-objective clustering technique based on the concepts of stability and symmetry, Springer, Knowl Inf Syst (2010) 23:1–27.

[27] S. Saha, S. Bandyopadhyay, A new multi-objective simulated annealing based clustering technique using symmetry, Pattern Recognition Letters 30 (2009) 1392–1403.

[28] S. Saha, S. Bandyopadhyay, A generalized automatic clustering algorithm in a multi-objective framework, Applied Soft Computing 13 (2013) 89–108.

[29] S. Bandyopadhyay and S. K. Pal, ``Classification and Learning Using Genetic Algorithms: Applications in

Bioinformatics and Web Intelligence", Springer, Heidelberg, 2007.

[30] S. K. Pal and S. Mitra, ``Fuzzy versions of Kohonen's net and MLP-based classification: Performance evaluation for certain non-convex decision regions", Information Sciences, Vol. 76, pp. 297-337, 1994.

[31] S. Mitra and S. K. Pal, ``Fuzzy multi-layer perceptron, inferencing and rule generation", IEEE Transactions on Neural Networks, Vol. 6, pp. 51-63, 1995.

[32] S. Bandyopadhyay and U. Maulik, ``An Evolutionary Technique Based on K-Means for Optimal Clustering in R^N", Information Sciences, vol. 146, pp. 221-237, 2002.

[33] S. Bandyopadhyay, S. Saha, 2007, GAPS: A Clustering Method Using A New Point Symmetry Based Distance Measure. Pattern Recog., 40, pp. 3430-3451.

[34] S. Bandyopadhyay, S. Saha,, 2008, A Point Symmetry Based Clustering Technique for Automatic Evolution of Clusters. IEEE Transactions on Knowledge and Data Engineering.

[35] S. Saha, S. Bandyopadhyay, automatic pixel classification in remote sensing satellite imagery using a new multi-objective simulated annealing based clustering technique, *Seminar on Spatial Information Retrieval, Analysis, Reasoning and Modelling 18th-20 th March 2009. ISI-DRTC, Bangalore, India*.

[36] U. Maulik and S. Bandyopadhyay, ``Genetic Algorithm-based Clustering Technique", Pattern Recognition, vol.32, pp. 1455-1465, 2000.

[37] W. Ma, L. Jiao, M. Gong, Immu-nodominance and clonal selection inspired multi-objective clustering, Progress in Natural Science 19 (2009) 751–758.

[38] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu., Understanding of internal clustering validation measures. In *IEEE ICDM*, pages 911–916, 2010.

[39] Z. Ansari, A.V. Babu, M.F. Azeem, W. Ahmed, Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions, World of Computer Science and Information Technology Journal (WCSIT), Vol. 1, No. 5, 217-226, 2011.