Volume 5, Issue 2, February 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Multi-objectives K-Modes Data Clustering Algorithm based on Self-Adaptive Differential Evolution

Omar S. Soliman, Doaa A. Saleh

Faculty of Computers and Information,

Cairo University, Egypt

Abstract— This paper proposes a Multi-objective K-Modes Data Clustering (MODEK-Modes) algorithm based on Self-Adaptive Differential Evolution (DE) for categorical data sets. The proposed algorithm purposes to improve the performance of data clustering through introducing three data clustering validity indexes. The proposed algorithm has three objectives: the first is a symmetry-index to maximize similarity within clusters; the second is a compactness index to maximize dissimilarity among clusters; and the last one is a validity Silhouette index to improve the validity of data clustering. The used validity indexes must be adjusted to be more appropriate for categorical data sets. Self-adaptive DE is similar to the traditional DE algorithm except two changes in the mutation and the crossover operations [31], where DE is a global optimization technique [15]. The proposed algorithm is implemented and evaluated using seven benchmark data sets obtained from the UCI Machine Learning Repository [5] and compared with different 4 data clustering algorithms that traditional K-mode, weighted K-mode [9], and significant k-mode algorithms [44], and MODEK-Modes by using several criteria (as ARI, Purity, Recall, Precision, Minkowski score, and Silhouette index). The experimental results showed that the proposed algorithm is performing well compared with the previous algorithms.

Keywords— K-modes, Multi-objective data clustering, NSGA-II, Self-Adaptive DE, validity indexes.

I. INTRODUCTION

The partitional clustering algorithms such as K-means are very efficient for processing large numeric datasets. But, the K-means clustering algorithm fails to handle datasets with categorical attributes because it minimizes the cost function by calculating means and distances. Furthermore, K-Modes clustering algorithm is developed to introduce a new dissimilarity measure to cluster categorical data [6], [14], [20], [25], [43], [49]. K-modes replaces means of clusters with modes (most frequent attribute value in an attribute), and uses a frequency based method to update modes in the clustering process to minimize the cost function [1], [10], [44]. In addition [2], K-Medoids algorithm, instead of computing the mean of feature vectors, selects cluster medoid for each cluster. The cluster medoid is defined as the most centrally located element in that cluster, i.e., it is the point from which the sum of the distances of the other points of the cluster is the minimum. Most of data clustering studies has been designed to deal with numerical or categorical data, but, in [17] and [49], they introduced a iterative clustering algorithm which have an ability to deal with categorical, numerical, and mixed attributes.

Validity clustering indexes are necessary to evaluate the performance of the tested data clustering algorithm [4]. Some recently studies used the validity indexes as objective functions in a multi-objective framework [39]. In multi-objective clustering, the goodness of data clustering should be judged not only by the clustering algorithm that generated it, but also by external and/or internal assessment criteria [33], [50]. Validity indexes [7], [22], [23], [24], [30], [33], [51] consist of 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 (EA) [13] are very successful in carrying out the optimization of multiple objectives. MOO deals with the real-world problems where there are several objectives that should be optimized simultaneously. In general, a MOO algorithm usually introduces a set of solutions that are not dominated by any solution. During recent years, many multi-objectives evolution algorithms, such as multi-objective EA (MOEA), have been suggested to solve the MOO problems. *Differential evolution* (DE) is considered a branch of evolutionary algorithms developed for optimization problems over continuous domains. DE is also considered as an extension 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 [19]. In [31], Self- adaptive DE algorithm is defined 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 (MODEK-Modes). The proposed algorithm is developed for categorical data sets. The reset of the paper is organized as follows: section 2 introduces related works of multi-objective data clustering, backgrounds for Self-Adaptive DE and the K-modes algorithm. Section 3 firstly presents the modifications of the used

validity indexes, then the proposed mathematical model, and finally the procedure of MODEK-Modes algorithm; Section 4 presents the used data sets, experimental results, discussion and analysis of obtained results; where the last section is devoted to the conclusion.

II. RELATED WORK & BACKGROUNDS

Clustering is considered 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 [2], the authors proposed 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 [12], multi-objective DE crisp clustering algorithm was introduced in 2010 for categorical data by also 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-objectives framework. DE data clustering algorithm was proposed with two objectives that the Xie-Beni index and Euclidean distances in [15]. In [47], the authors selected two complementary objectives that compactness and connectedness of clusters based on AIS.

Symmetry-index and Euclidean distances are used as objective functions and solved by SA in [42]. 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 [16]. In [17], the authors proposed a Multi-objective C-means data clustering selected using used SA, these objectives was symmetry-index, connectively-index, and I-index The Xie-Beni index *XBq* and a penalized version was selected as the two objectives based on DE and the FCM function *Jq* [19]. In [40] and [37], symmetry- index and average of symmetry-index was used as objective functions for achieving stability among clusters. In [3], the authors selected two objectives that the Xie-Beni index and Euclidean distances based on FPSO. Multi-objectives data clustering algorithm was proposed in [38] 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).

A. SELF-ADAPTIVE DIFFERENTIAL EVOLUTION

DE is a population-based global optimization algorithm that uses a real-coded representation [13]. 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 [19]. Self-adaptive DE algorithm is the same as the traditional DE algorithm except two changes in the mutation and the crossover operations:

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:

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

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

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

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

B. K-MODES ALGORITHM

The K-modes algorithm extends the K-means paradigm to cluster categorical data by removing the barrier imposed by K-means through following modifications [44]:

1. Using a simple matching dissimilarity measure or the Hamming distance for categorical data objects.
2. Replacing means of clusters by their modes (cluster centers).

To describe K-Modes algorithm [14]: let D = {x_1, x_2, …, x_n} be a categorical data set with n objects each of which is described be *d* categorical attributes A_1, A_2, …., A_d. Attribute A_j ($1 \le l \le d$) has n_j categories, i.e., DOM $(A_j) = \{a_{l1}, a_{l2}, \dots, a_{lnl}\}$. Let the cluster center be represented by $z_j = \{z_{j1}, z_{j2}, \dots, z_{jd}\}$ *for* $1 \le j \le k$, where *k* is the number of clusters. Simple matching [8], [9] is defined as a common approach in which comparison of two identical categorical values yields a difference of zero while comparison of two distinct categorical values yields a difference of one. The simple matching distance measure between x and y in D is mathematically defined as:

$$d_C(x, y) = \sum_{l=1}^{d} \delta(x_l, y_l) \quad (3)$$

where x_j and y_j are the j^{th} components of x and y, respectively, and

$$\delta(x_j, y_j) = \begin{cases} 0 & if\ x_l = y_l \\ 1 & if\ otherwise \end{cases}$$

To update the centroid of clusters by $z_{jl} = a_{lr} \in DOM(A_l)$ (4)

Where $r = \arg max_{1 \le t \le n_j} [\sum_{i,x_{il}=a_{lt}} \mu_{ji}^{\alpha}]$, and

$\sum_{i,x_{il}=a_{lr}} \mu_{ji}^{\alpha} \geq \sum_{i,x_{il}=a_{lt}} \mu_{ji}^{\alpha}$, $1 \le t \le n_j$ $for\ 1 \le l \le d$ and $1 \le j \le k$

III. THE PROPOSED ALGORITHM

The proposed multi-objectives (MODEK-Modes) algorithm uses three of validity data clustering indexes as objective functions. The proposed algorithm is developed for categorical data sets, so that, the used validity indexes have to be modified to be more adequate for the nature of categorical data. This algorithm is solved by using Self-Adaptive DE. Furthermore, this section is divided into three sub-sections that: 1) the modifications which operated on the used validity indexes, 2) the mathematical model, and 3) the procedure of MODEK-Modes algorithm.

A. Modified Validity Indexes

The most validity indexes are mainly processing based on Euclidian distances (means of clusters). But, k-modes algorithm is working on mode of centers. The used validity indexes are that symmetry index, compactness index, and silhouette validity index. Next, the modified validity indexes are introduced in details, as follows:

[1] Modified symmetry index (*MSym-index*): The modified symmetry index is considered as an extension of the original symmetry index [41]. The new validity index has two adjustments in that:

- Replacing Euclidian distances with matching distances.
- Determining the reflected point.

Let a point $\bar{x}$, the symmetrical (reflected) point of $\bar{x}$ with respect to a particular center $\bar{c}$ will be calculated by this formula $2 \times d_m(\bar{x}, \bar{c})$, where $d_m(\bar{x}, \bar{c})$ is calculated by simple matching distance by using Eq. (2) [32]. Let the first and the second unique nearest neighbors to $\bar{x}^*$ be at simple matching distance of d_1 and d_2, respectively. Then $d_{ps}(\bar{x}, \bar{c}) = \frac{d_1 + d_2}{2} \times d_m(\bar{x}, \bar{c})$, where $d_m(\bar{x}, \bar{c})$ is the simple matching distance between the point $\bar{x}$ and $\bar{c}$. Then, *MSym-index* is a new cluster validity function for categorical data sets which measures the overall average symmetry with respect to the cluster centers. The new cluster validity function for categorical data sets is defined as:

$$Maximize \qquad MSym(k) = \left(\frac{1}{k} \times \frac{1}{\varepsilon_k} \times D_k\right) \qquad (5)$$

$$\varepsilon_k = \sum_{i=1}^{k} E_i \ , E_i = \sum_{j=1}^{n_i} d_{ps}^{*}(\bar{x}_j^i, \bar{c}_i), and \ D_K = max_{i,j=1}^{k} \|\bar{c}_i - \bar{c}_j\|$$

where, D_K is the maximum matching distance between two cluster centers among all pairs of centers.

[2] Modified connectivity based cluster validity index (*MCon-index*): The new cluster validity index is also considered as an extension of the compactness index. The compactness index [46] is able to detect the appropriate partitioning from data sets having clusters of any shape, size or convexity as long as they are well separated. Furthermore, this index mainly depends on measuring the distances in more than one form, so that, the adjustment in this index is by

- Replacing Euclidian distances with matching distances which used in all terms of this index.

Suppose the clusters formed are by C_K, for $k= 1, \ldots, K$, where K is the number of clusters. Then the medoid of the Kth cluster, denoted by $\overline{m_k}$, is the point of that cluster which has the minimum average distance to all the other points in that cluster. $\bar{x}_{minidex}^{k} = argmin_{i=1}^{n_k} \frac{\sum_{j=1}^{n_k} d_m(x_i^k, x_j^k)}{n_k}$, where $d_m(\bar{x}_i^k, \bar{x}_j^k)$ is calculated by the matching distance, n_k is the total number of data points in the cluster kth, $\bar{x}_i^k$ refers to the data point ith in the cluster kth, then $\bar{m}_k = \bar{x}_{minidex}^{k}$. Then the MCon-index function will be as follows:

$$Minimize \qquad MCon = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_k} d_{short}(\bar{m}_i, \bar{x}_j^i)}{n \times min_{i,j=1 \wedge i \neq j}^{k} d_{short}(\bar{m}_i, \bar{m}_j)} \qquad (6)$$

where $d_{short}(\bar{m}_i, \bar{x}_j^i)$ & $d_{short}(\bar{m}_i, \bar{m}_j)$ will be also calculated by the matching distance. The smaller values of *MCon-index* correspond to good partitioning. Furthermore, achieving the good partitioning is through the minimum value of the *MCon-index*.

[3] *Modified Silhouette Validity Index* (*MSil-Index*): Silhouette Validity Index is considered from the internal validity indexes.

- *MSil-Index* is also considered as an extension of silhouette index by replacing Euclidian distances with matching distances in calculating silhouettes width. The silhouettes width of ith data point is computed by following this formula

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

where *a(i)* is average dissimilarity of ith data point to all other points in the same cluster by the matching distances; 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.

$$Maximize \qquad MSil(k) = \frac{1}{N}\sum_{i=1}^{N} S_i \qquad (7)$$

B. The Mathematical Model

The proposed algorithm has three objective functions in addition to three constrain as shown in fig. 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.

$$\textbf{Maximize} \qquad MOP = \left(MSym(K), \frac{1}{MCon(K)}, MSil(K)\right) \qquad (8)$$

Subject to

$$\sum_{j=1}^{n}\mu_{kj} \geq 1 \quad for\ k = 1, \dots, K \qquad (9),$$

$$\sum_{k=1}^{K}\mu_{kj} = 1 \quad for\ j = 1, \dots, n \qquad (10),$$

$$\sum_{k=1}^{K}\sum_{j=1}^{n}\mu_{kj} = n \qquad (11)$$

Fig. 1 The mathematical model of the proposed MODEK-Modes algorithm

In Eq. (8), *MSym*, *MCon*, and *MSil* refer to modified symmetry-index, modified compactness index, and modified validity Silhouette index, with respectively. With respect to the constraints (Eq. 9, 10, and 11), *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 clustering, $\mu_{kj} = 1\ if\ x_j \in C_k$, otherwise $\mu_{kj} = 0$.

C. The proposed MODEK-Modes Algorithm

The proposed MODEK-Modes 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 the inputs and the output of this algorithm; steps (3 and 4) are an initialization steps for some required parameters, the first population, and the center of clusters; step 5 evaluates each individual in the population, where each individual has three values according to the three objective functions; rest of the steps are considered the main body of the proposed algorithm which start with calculating the modes of centers matrix for each individual by the matching distances then updating distances matrixes, 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 fig. 2.

1: ***Input:*** *D = Categorical Dataset, N = Data objects, M = Number of attributes in the data, K = Number of data clusters, and Max= Maximum number of iterations.*
2: ***Output:*** *dividing D data set on* ***K*** *clusters with high efficiency and performance****.***
3: *Initialize parameters and the first population generation of n individuals.*
4: *Initialize centers of clusters.*
5: *Evaluate Fitness evaluation of individuals.*
6: *While Stopping Criterion not met; do,*
   6.1. ***For*** *each individual Pi (i = 1. . . NP) from P* ***do****:*
      a) Update modes of clusters by using Eq. (4) for each individual in the population.
      b) Update distance between data objects and the new centers of clusters by ***the matching distance*** *by using Eq. (3).*
      c) Calculate distance among clusters by ***the matching distance*** *by using Eq. (3).*
      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.*
   6.2. If the population has more than pop-Size individuals, truncate it
   6.3. Randomly enumerate the individuals in P.

6.4. If not met stopping criterion, go to step (6.1)
7: ***End For***
8: ***End While***
9: Determine a set of non-dominate solutions (individuals) from the new-population.
10: Select optimal solution from the set of non-dominate solutions according to ARI measure performance.

Fig. 2 MODEK-Modes 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], [35], [36] measure performance. Next section will discussed used data sets, obtained results, and analysis under an umbrella of experimental results.

IV. EXPERIMENTAL RESULTS

The proposed algorithm is compared with four different algorithms traditional K-Modes, WK-Modes, SK-Modes, and MOGAK-Modes algorithms. Firstly, for more experimental results, MOGAK-Modes algorithm is developed based on the mathematical model (in fig. 1) and NSGA-II by helping these references that [11], [18], [26], [27], [28], [29], [34], and [48]. The parameters of MOGAK-Modes algorithm are as follows: (maximum no. of iterations) Gmax = 50, pop = 100, $\beta = 0.1$ (used in selection operator), and pm= 0.01 (is the mutation parameter). Secondly, the proposed MODEK-Modes algorithm is developed based on Self-Adaptive DE to get better results, where DE is a global optimization technique.

A. Data Sets

For implementing the proposed algorithm, we use the seven standard data sets (shown in Table 1) obtained from the UCI Machine Learning Repository [5] by using VC++ to test the performance of the proposed algorithm. These data sets are introduced as follows:

1) **Small soybean data**: the data set has 47 records, each of which is described by 35 attributes.
2) **Mushroom data**: it consists of 8124 data objects and 22 categorical attributes.
3) **Dermatology data**: it consists of 366 data objects and 33 attributes.
4) **Lung-Cancer data**: this data set consists of 32 data objects and 56 attributes.
5) **Breast-Cancer data**: it consists of 699 data objects and 9 attributes.
6) **Zoo data**: it consists of 101 data objects and 17 attributes.
7) **Congressional vote data**: It consists of 435 data objects and 16 categorical attributes.

B. AC, PR, and RE Results

The obtained results of the 100 independent runs are summarized and tabulated in tables from 1 and 2. Table 1 contains the best result of (AC, PR, and RE) [45] in the 100 runs and the computed rank (the numbers in between brackets). The proposed algorithm is compared with four different algorithms; the first is the traditional k-modes algorithm; the second is the weighted k-modes algorithm [10], and the last one is significant k-mode algorithm [39]. The rank is also computed according to the values of each performance measure. This rank is taking values from 1 to 4, where the best will get rank with value one and the worst will take four.

Table 1 Best of AC, PR, and RE for 100 runs of five algorithms on the seven data sets

		K-Modes	*WK-Modes*	*SK-Modes*	*MOGAK-Modes*	*MODEK-Modes*
Soybean	**AC**	0.8553 (4)	0.8613 (3)	0.6809 (5)	0.8926 (2)	0.9181 (1)
	PR	0.9020 (3)	0.8948 (4)	0.7549 (5)	0.9028 (2)	0.9373 (1)
	RE	0.8407 (4)	0.8471 (3)	0.7176 (5)	0.8652 (1)	0.8599 (2)
Mushroom	**AC**	0.7176 (3)	0.7106 (4)	0.5086 (5)	0.7681 (2)	0.7975 (1)
	PR	0.7453 (3)	0.7414 (4)	0.7303 (5)	0.7911 (1)	0.7811 (2)
	RE	0.7132 (3)	0.7056 (4)	0.5256 (5)	0.7421 (2)	0.7436 (1)
Dermatology	**AC**	0.6869 (3)	0.6854 (4)	0.6502 (5)	0.7052 (2)	0.7274 (1)
	PR	0.7629 (4)	0.7692 (3)	0.5601 (5)	0.7855 (2)	0.7898 (1)
	RE	0.5750 (4)	0.5765 (3)	0.5512 (5)	0.6472 (2)	0.6577 (1)
Lung-Cancer	**AC**	0.5313 (4)	0.5497 (3)	0.4688 (5)	0.6177 (1)	0.6031 (2)
	PR	0.5880 (4)	0.5965 (3)	0.5079 (5)	0.6352 (1)	0.6311 (2)
	RE	0.5374 (4)	0.5626 (3)	0.4838 (5)	0.6081 (1)	0.5952 (2)
Breast-Cancer	**AC**	0.8482 (5)	0.8530 (4)	0.9127 (2)	0.9043 (3)	0.9133 (1)
	PR	0.8731 (5)	0.8733 (4)	0.9318 (1)	0.9279 (3)	0.9280 (2)
	RE	0.7893 (5)	0.7968 (4)	0.8783 (1)	0.8692 (3)	0.8774 (2)

Zoo	**AC**	0.8011 (4)	0.9106 (2)	0.8911 (3)	0.9264 (1)	0.9638 (1)
	PR	0.7671 (4)	0.8101 (3)	0.7224 (5)	0.8323 (2)	0.8377 (1)
	RE	0.7661 (2)	0.7482 (4)	0.7398 (5)	0.7638 (3)	0.7984 (1)
Cong. Vote	**AC**	0.8352 (5)	0.8620 (3)	0.8506 (4)	0.8737 (2)	0.8811 (1)
	PR	0.7962 (5)	0.8811 (3)	0.8484 (4)	0.8934 (2)	0.8998 (1)
	RE	0.8117 (5)	0.8792 (2)	0.8672 (4)	0.8763 (3)	0.8844 (1)
Rank Average		*4*	*3.24*	*4.143*	*1.86*	*1.33*

The proposed MODEK-Modes algorithm gets the minimum value of the average rank to be 1.133. MOGAK-Modes algorithm obtains on the second place in the average rank to be 1.86, then WK-Modes gets 3.24, and finally traditional K-Modes and SK-Modes algorithms obtained on 4 and 4.143, with respectively. Table 2 introduces the mean and standard division of AC, PR, and RE for 100 independent runs. The proposed algorithm obtains the best results in the most data sets according to the average values.

Table 2 Means and standard deviation of AC, PR, and RE for 100 runs of five algorithms on the seven data sets

		K-Modes	***WK-Modes***	***SK-Modes***	***MOGAK-Modes***	***MODEK-Modes***
Soybean	AC	0.8011± 0.23	0.8382± 0.17	0.6316± 0.14	0.8621± 0.13	0.8911± 0.01
	PR	0.8572± 0.11	0.8723± 0.03	0.7362± 0.01	0.8951± 0.01	0.9037± 0.03
	RE	0.8117± 0.3	0.8202± 0.2	0.7009± 0.02	0.8371± 0.01	0.8462± 0.11
Mushroom	AC	0.7021± 0.02	0.7011± 0.11	0.4966± 0.3	0.7407± 0.2	0.7822± 0.17
	PR	0.7381± 0.02	0.7325± 0.07	0.7011± 0.13	0.7401± 0.01	0.7388± 0.21
	RE	0.7069± 0.14	0.6942± 0.21	0.5002± 0.14	0.6911± 0.11	0.7142± 0.13
Dermatology	AC	0.6398± 0.07	0.6421± 0.04	0.6382± 0.09	0.6872± 0.02	0.7081± 0.01
	PR	0.7422± 0.11	0.7206± 0.2	0.5411± 0.13	0.7419± 0.01	0.7677± 0.07
	RE	0.5477± 0.03	0.5557± 0.11	0.5322± 0.2	0.6124± 0.12	0.6383± 0.08
Lung-Cancer	AC	0.5198± 0.08	0.5155± 0.01	0.4399± 0.08	0.5732± 0.07	0.5982± 0.11
	PR	0.5569± 0.12	0.5491± 0.3	0.4911± 0.02	0.5543± 0.12	0.5493± 0.1
	RE	0.5189± 0.09	0.5458± 0.08	0.4538± 0.13	0.5792± 0.3	0.5622± 0.17
Breast-Cancer	AC	0.8278± 0.06	0.8365± 0.21	0.8973± 0.11	0.8881± 0.09	0.9088± 0.12
	PR	0.8611± 0.05	0.8659± 0.01	0.9011± 0.11	0.8953± 0.03	0.9206± 0.08
	RE	0.7689± 0.13	0.7788± 0.11	0.8432± 0.09	0.8611± 0.13	0.8852± 0.03
Zoo	AC	0.8976± 0.2	0.9043± 0.01	0.8736± 0.3	0.9038± 0.05	0.9322± 0.07
	PR	0.7439± 0.11	0.7965± 0.08	0.7062± 0.3	0.8077± 0.2	0.8293± 0.02
	RE	0.7661± 0.02	0.7482± 0.07	0.7398± 0.06	0.7638± 0.03	0.7624± 0.11
Cong. Vote	AC	0.8169± 0.01	0.8558± 0.2	0.8128± 0.05	0.8562± 0.01	0.8791± 0.04
	PR	0.7734± 0.06	0.8741± 0.05	0.8332± 0.11	0.8712± 0.04	0.8823± 0.01
	RE	0.7929± 0.11	0.8617± 0.2	0.8429± 0.3	0.8621± 0.2	0.8755± 0.02

C. Measures Performance

Firstly, Adjust Rand Index (ARI) is considered the external clustering validity index [35], [36]. Table 3 represents the best values, mean, and standard division of ARI through 100 independent runs. The proposed algorithm is compared with the previous algorithms. The proposed algorithm also achieved the best results of ARI for the most real data sets.

Table 3 ARI values for the proposed algorithm & the previous algorithms on the seven data sets.

		K-Modes	***WK-Modes***	***SK-Modes***	***MOGAK-Modes***	***MODEK-Modes***
Soybean	***Best***	0.7411	0.8191	0.6632	**0.8538**	**0.9284**
	Mean	0.7293± 0.02	0.8054 ± 0.05	0.7411± 0.01	**0.8328** ± 0.01	0.9011± 0.01
Mushroom	***Best***	0.3682	0.3811	0.3911	**0.4102**	**0.6482**
	Mean	0.3500 ± 0.06	0.3586 ± 0.11	0.3472 ± 0.07	**0.3711**± 0.02	0.5982± 0.03
Dermatology	***Best***	0.5013	0.4920	0.5002	**0.6092**	**0.6321**
	Mean	0.4831 ± 0.05	0.4801 ± 0.09	0.4729 ± 0.054	**0.5741** ± 0.04	0.5837± 0.01
Lung-Cancer	***Best***	0.6252	**0.6488**	0.6102	0.6431	0.6642
	Mean	0.6110 ± 0.11	0.6366 ± 0.06	0.6022 ± 0.21	**0.6374** ± 0.05	0.6392± 0.07
Breast-	***Best***	0.4938	**0.5302**	0.4992	0.5298	0.5452

Cancer	***Mean***	0.4828 ± 0.09	**0.5130** ± 0.04	0.4791 ± 0.19	0.5111 ± 0.01	0.5348± 0.09
Zoo	***Best***	0.6963	0.7292	0.6992	**0.7433**	**0.7611**
	Mean	0.6722 ± 0.02	0.7141 ± 0.13	0.6681± 0.12	**0.7283** ± 0.01	0.7353± 0.02
Cong. Vote	***Best***	0.5741	0.5745	0.5522	**0.5808**	**0.5945**
	Mean	0.5313± 0.11	0.5345 ± 0.07	0.5155± 0.01	**0.5688** ± 0.06	0.5711± 0.04

Secondly, The *Minkowski score* [12] is the normalized distance between the two matrices. Lower *Minkowski score* implies better clustering solution, and a perfect solution will have a score zero. From fig. 3, the proposed algorithm mostly achieved the minimum values of the *Minkowski score*, where the minimum value of this score is the best performance.

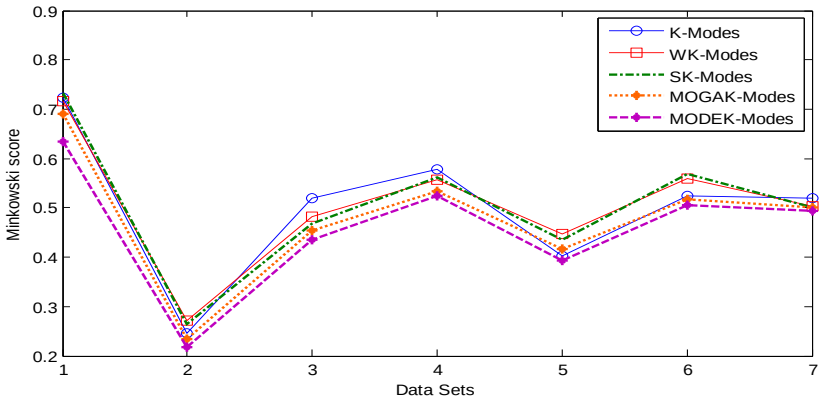

Fig. 3 The Minkowski score for the previous five clustering algorithms

Finally, on the other hand, the proposed data clustering algorithm need to be internally evaluated by one of the internal validity index like that the Silhouette index. Actually, this evaluation is done between the proposed algorithm and the previous algorithms (see, fig. 4). Whenever the value of Silhouette index is close to one, then this clustering is well.

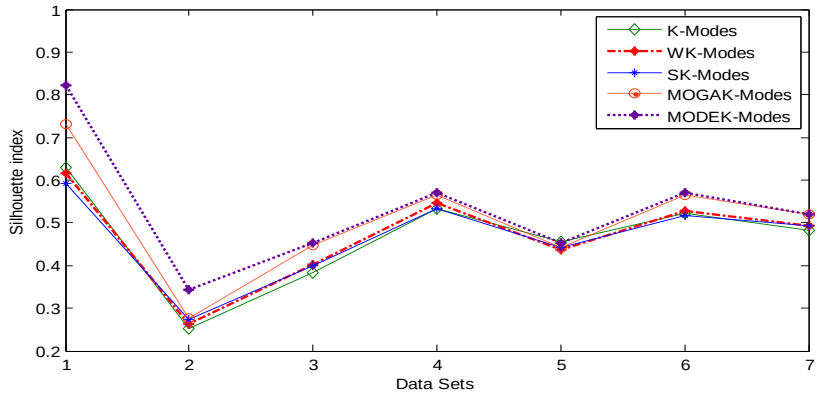

Fig. 4 The Silhouette index for the previous five clustering algorithms

V. CONCLUSION

This paper introduced an algorithm for improving the performance of data clustering through achieving 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 the modified symmetry-index to maximize similarity within clusters, the modified compactness index to maximize dissimilarity among clusters, and the modified Silhouette index to improve the validity of data clustering. The used validity indexes were adjusted to be more appropriate for categorical data sets. The proposed algorithm was implemented on seven benchmark data sets and compared with four different data clustering algorithms traditional K-mode, weighted K-mode, and significant k-mode algorithms, and MOGAK-Modes by using several criteria (as ARI, Purity, Recall, Precision, Minkowski score, and Silhouette index). The obtained results showed that the proposed MODEK_Modes algorithm performed well compared with its compared algorithms.

REFERENCES

[1] A. Ammar, Z. Elouedi, P. Lingras, K-Modes Clustering Using Possibilistic Membership, IPMU 2012, Part III, CCIS 299, pp. 596–605, 2012.

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

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

[4] 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*.

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

[6] D. Ienco, R. G. Pensa, R. Meo, From Context to Distance: Learning Dissimilarity for Categorical Data Clustering, ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 6 Issue 1, March 2012.

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

[8] F. Cao, J. Liang, D. Li, L. Bai, Ch. Dang, A dissimilarity measure for the k-Modes clustering algorithm, Knowledge-Based Systems 26 (2012) 120–127.

[9] F. Cao, J. Liang, D. Li, X. Zhao, A weighting k-modes algorithm for subspace clustering of categorical data, Neurocomputing 108 (2013) 23–30.

[10] G. David, A. Averbuch, SpectralCAT: Categorical spectral clustering of numerical and nominal data, Pattern Recognition 45 (2012) 416–433.

[11] H. Ghiasi, D. Pasini, L. Lessard, A non-dominated sorting hybrid algorithm for multi-objective optimization of engineering problems, Engineering Optimization, Vol. 43, No. 1, January 2011, 39–59.

[12] 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.

[13] 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.

[14] I. Saha, J. P. Sarkar, U. Maulik, Rough Set Based Fuzzy K-Modes for Categorical Data, SEMCCO 2012, LNCS 7677, pp. 323–330, 2012.

[15] 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.

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

[17] J. Ji, T. Bai, Ch. Zhou, Ch. Maa, Z. Wang, An improved k prototypes clustering algorithm for mixed numeric and categorical, Neurocomputing 120 (2013) 590–596.

[18] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation, Vol. 6, No. 2, April 2002.

[19] 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 (2009) 1424-8220.

[20] L. Baia, J. Lianga, Ch. Dang, F. Cao, A novel fuzzy clustering algorithm with between-cluster information for categorical data, Fuzzy Sets and Systems 215 (2013) 55–73.

[21] L. Hubert, P. Arabie, (1985) Comparing partitions. Journal of Classification, 193–218.

[22] 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.

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

[24] 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.

[25] M. H. C. Law, A. P. Topchy, A. K. Jain, Multiobjective Data Clustering, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004.

[26] M. R. Rani, H. Selamat, H. Zamzuri, Z. Ibrahim, Multi-Objective Optimization for Pid Controller Tuning Using the Global Ranking Genetic Algorithm, International Journal of Innovative Computing, Information and Control, Vol. (8), No. (1A), Jan. 2012.

[27] M. Zarabian, S. T. A. Niaki, M. S. Mehrabad, A NSGA-II algorithm to solve a bi-objective optimization of the redundancy allocation problem for series-parallel systems, 2012 2nd International Conference on Industrial Technology and Management (ICITM 2012).

[28] N. Chase, M. Rademacher, E. Goodman, R. Averill, R. Sidhu, A Benchmark Study of Multi-Objective Optimization Methods, BMK-3021 Rev. 06.09.

[29] N. Srinivas, K. Deb, Multi-objective Optimization Using Non-dominated Sorting in Genetic Algorithms, Evolutionary Computation, 1994, Vol. 2, No. 3, pages 221-248.

[30] 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.

[31] 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.

[32] O. S. Soliman, D. A. Saleh, S. Rashwan, A Bio Inspired Fuzzy K-Modes Clustring Algorithm, Springer Journal, ICONIP 2012, Part III, LNCS 7605, pp. 663 – 669.

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

[34] R. G. L. D'Souza, K. Ch. Sekaran, A. Kandasamy, Improved NSGA-II Based on a Novel Ranking Scheme, Journal of Computing, Volume 2, Issue 2, February 2010, ISSN 2151-9617.

[35] 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.

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

[37] S. Bandyopadhyay, Multiobjective 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.

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

[39] S. Saha, A. Eknal, K. Gupta and S. Bandyopadhyay, Gene Expression Data Clustering Using A Multiobjective Symmetry Based Clustering Technique. Computers in Biology and Medicine, Volume 43, Issue 11, Pages 1965-1977.

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

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

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

[43] S. W. Purnami, J. M. Zain, A. Embong, Reduced Support Vector Machine Based on k-Mode Clustering for Classification Large Categorical Dataset, ICSECS 2011, Part II, CCIS 180, pp. 694–702, 2011.

[44] Sh. S. Khan, A. Ahmed, Cluster center initialization algorithm for K-modes clustering, Expert Systems with Applications 40 (2013) 7444–7456.

[45] T. A. Al-Fayyaz, Optimization of multi-objective reservoir operation system, Master Thesis, Department of Civil Engineering National University of Singapore, 2004.

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

[47] W. Ma, L. Jiao, M. Gong, Immunodominance and clonal selection inspired multiobjective clustering, Progress in Natural Science 19 (2009) 751–758.

[48] Y. Zhang, M. Harman, S. A. Mansouri, The Multi-Objective Next Release Problem, GECCO'07, July 7–11, 2007, London, England, United Kingdom.

[49] Y. Ming Cheung, H. Jia, Categorical-and-numerical attribute data clustering based on aunified similarity metric without knowing cluster number, Pattern Recognition 46 (2013) 2228–2238.

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

[51] 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.

