Dear Author

Here are the proofs of your article.

- You can submit your corrections online, via e-mail or by fax.
- For online submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and email the annotated PDF.
- For fax submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the journal title, article number, and your name when sending your response via e-mail or fax.
- Check the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- Check the questions that may have arisen during copy editing and insert your answers/corrections.
- Check that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the Edited manuscript.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please do not make changes that involve only matters of style. We have generally introduced forms that follow the journal’s style.
- Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections within 48 hours, we will send you a reminder.
- Your article will be published Online First approximately one week after receipt of your corrected proofs. This is the official first publication citable with the DOI. Further changes are, therefore, not possible.
- The printed version will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL:

http://dx.doi.org/10.1007/s10489-018-1350-8

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to:


Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.
<table>
<thead>
<tr>
<th>Article Title</th>
<th>APIN_R1: sophisticated SOM based genetic operators in multi-objective clustering framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Sub-Title</td>
<td></td>
</tr>
<tr>
<td>Article Copyright - Year</td>
<td>Springer Science+Business Media, LLC, part of Springer Nature 2018 (This will be the copyright line in the final PDF)</td>
</tr>
<tr>
<td>Journal Name</td>
<td>Applied Intelligence</td>
</tr>
<tr>
<td>Corresponding Author</td>
<td>Family Name: Saini, Particle: Naveen</td>
</tr>
<tr>
<td></td>
<td>Given Name: Saini, Suffix:</td>
</tr>
<tr>
<td></td>
<td>Organization: Indian Institute of Technology, Division: Department of Computer Science and Engineering, Address: Patna, Bihar, India, e-mail: <a href="mailto:naveen.pcs16@iitp.ac.in">naveen.pcs16@iitp.ac.in</a></td>
</tr>
<tr>
<td>Author</td>
<td>Family Name: Saha, Particle:</td>
</tr>
<tr>
<td></td>
<td>Given Name: Sriparna, Suffix:</td>
</tr>
<tr>
<td></td>
<td>Organization: Indian Institute of Technology, Division: Department of Computer Science and Engineering, Address: Patna, Bihar, India, e-mail: <a href="mailto:sripama@iitp.ac.in">sripama@iitp.ac.in</a></td>
</tr>
<tr>
<td>Author</td>
<td>Family Name: Harsh, Particle:</td>
</tr>
<tr>
<td></td>
<td>Given Name: Aditya, Suffix:</td>
</tr>
<tr>
<td></td>
<td>Organization: UPES, Division:</td>
</tr>
<tr>
<td></td>
<td>Address: Dehradun, India, e-mail: None</td>
</tr>
<tr>
<td>Author</td>
<td>Family Name: Bhattacharyya, Particle:</td>
</tr>
<tr>
<td></td>
<td>Given Name: Pushpak, Suffix:</td>
</tr>
<tr>
<td></td>
<td>Organization: Indian Institute of Technology, Division: Department of Computer Science and Engineering, Address: Patna, Bihar, India</td>
</tr>
</tbody>
</table>
Multi-objective clustering refers to the partitioning of a given collection of objects into various $K$-groups based on some similarity/dissimilarity criteria while optimizing different partition quality measures simultaneously. The current paper proposes an automated decomposition based multi-objective clustering technique, SOMDEA$_\text{clust}$, which is a fusion of self-organizing map (SOM) and multi-objective differential evolution. A novel reproduction operator is designed where the ensemble of multiple neighborhoods extracted using self-organizing map is used for constructing the variable mating pool size. The probabilities of selecting different sizes of the neighborhood are updated based on their performances in generating new improved solutions in the last few generations. Decomposition based selection scheme is also utilized in our paper which divides the multi-objective optimization (MOO) problem into a number of single objective subproblems. The objective functions corresponding to these subproblems are optimized in a collaborative manner by the use of MOO. The potentiality of the proposed framework is shown for clustering four real-life data sets and five artificial data sets in comparison to some existing multi-objective based clustering techniques, namely MOCK, SMEA$_\text{clust}$, MEA$_\text{clust}$, a single objective based genetic clustering technique, SOGA and a traditional clustering technique, K-means. To show the utility of SOM based reproduction operators, another decomposition based multi-objective clustering technique (MDEA$_\text{clust}$) without the use of SOM based operators is also developed in this paper. In order to show the efficacy of the proposed clustering technique in handling large data sets, two large scale datasets having more than 5000 data points are also utilized. As a real-life application, the proposed clustering technique is applied for scientific/web document clustering where a set of scientific/web documents are partitioned based on their content-similarities. Semantic representation is utilized to covert the text document into a real vector. Experimental results clearly illustrate the effectiveness of fusion of SOM and DE in developing an effective clustering technique.

Keywords
Clustering - Cluster validity indices - Self organizing map (SOM) - Differential evolutionary algorithm (DE) - Polynomial mutation - Multi-objective optimization (MOO)
APIN_R1: sophisticated SOM based genetic operators in multi-objective clustering framework

Naveen Saini1 · Sriparna Saha1 · Aditya Harsh2 · Pushpak Bhattacharyya1

Abstract

Multi-objective clustering refers to the partitioning of a given collection of objects into various $K$-groups based on some similarity/dissimilarity criteria while optimizing different partition quality measures simultaneously. The current paper proposes an automated decomposition based multi-objective clustering technique, SOMDEA_clust, which is a fusion of self-organizing map (SOM) and multi-objective differential evolution. A novel reproduction operator is designed where the ensemble of multiple neighborhoods extracted using self-organizing map is used for constructing the variable mating pool size. The probabilities of selecting different sizes of the neighborhood are updated based on their performances in generating new improved solutions in the last few generations. Decomposition based selection scheme is also utilized in our paper which divides the multi-objective optimization (MOO) problem into a number of single objective subproblems. The objective functions corresponding to these subproblems are optimized in a collaborative manner by the use of MOO. The potentiality of the proposed framework is shown for clustering four real-life data sets and five artificial data sets in comparison to some existing multi-objective based clustering techniques, namely MOCK, SMEA_clust, MEA_clust, a single objective based genetic clustering technique, SOGA and a traditional clustering technique, K-means. To show the utility of SOM based reproduction operators, another decomposition based multi-objective clustering technique (MDEA_clust) without the use of SOM based operators is also developed in this paper. In order to show the efficacy of the proposed clustering technique in handling large data sets, two large scale datasets having more than 5000 data points are also utilized. As a real-life application, the proposed clustering technique is applied for scientific/web document clustering where a set of scientific/web documents are partitioned based on their content-similarities. Semantic representation is utilized to covert the text document into a real vector. Experimental results clearly illustrate the effectiveness of fusion of SOM and DE in developing an effective clustering technique.

Keywords Clustering · Cluster validity indices · Self organizing map (SOM) · Differential evolutionary algorithm (DE) · Polynomial mutation · Multi-objective optimization (MOO)

1 Introduction

1.1 Background

Clustering [17], also known as unsupervised learning, is a technique which divides the input space into $K$-partitions based on some dissimilarity/similarity criteria where $K$ may or may not be known beforehand. Traditional clustering techniques like K-means [17] etc., in general optimize a single cluster quality measure and capture different cluster characteristics like compactness [23], separation, connectivity [28], symmetricity or density [27] to find the appropriate partitioning. In order to automatically detect the number of clusters present in a dataset, these algorithms are executed with varying number of clusters, $K$, and the values of a cluster validity index [23] are calculated for these partitionings. Finally the partitioning corresponding to the optimal value of the cluster validity index is selected as the optimal one. But most of the real-life data sets contain clusters having different geometrical shapes. By optimizing a single cluster quality measure, it is difficult to detect clusters having different shapes, structures and densities. Moreover the value of $K$ may not be known beforehand for many real-life data sets. Therefore it is necessary to develop some...
multi-objective [11] clustering techniques [29] which can automatically partition the data set to detect clusters having different structures and shapes, by optimizing more than one cluster validity measures (indices), simultaneously.

1.2 Related works

In recent years, many multi-objective clustering techniques are developed utilizing different meta-heuristic techniques like particle swarm optimization [19], genetic algorithms (GA) [2], differential evolution (DE) [33] etc. In [2], GCUK, an automatic clustering approach was proposed optimizing single cluster validity measure, Xie-Beni index [23] and this approach is able to detect only hyperspherical shaped clusters. In [5], a symmetry distance based genetic clustering algorithm is proposed namely VGAPS clustering, which evolves proper partitioning as well as the number of clusters automatically from a data set. But it also optimizes a single cluster validity measure, symmetry distance based Sym index [3] and is able to detect only symmetrical shaped clusters. Both [2] and [5] are rarely applicable to different types of data sets with different properties.

To overcome these limitations, in recent years [27, 29], some symmetry based automatic multi-objective clustering techniques utilizing archived multi-objective simulated annealing [6] process as the underlying optimization technique are proposed. Handl et al. [15] proposed an automatic multi-objective clustering technique, MOCK, optimizing two objective functions simultaneously. Main limitation of MOCK is that it can determine only some hyper-spherical shaped or well-separated clusters and can not detect overlapping clusters. In addition, complexity of MOCK increases linearly as number of data points increases.

Some automatic multi-objective clustering techniques using differential evolution algorithm [26] as the underlying optimization strategy are proposed in [8, 33] and they have shown that differential evolution has faster convergence rate as compared to other evolutionary algorithms like NSGA-II [13] and can serve as a better optimization strategy for multi-objective clustering techniques. In [30], a SOM based automatic clustering algorithm is proposed using multi-objective differential evolution (MODE) algorithm [26] as the underlying optimization strategy. Here some SOM based genetic operators were proposed in the MODE based framework. The properties of SOM are used for constructing the mating pool which is then utilized for developing the genetic operators. But this approach suffers from the use of fixed mating pool size. All the above discussed clustering techniques are automatic in nature, but their applications are shown only for partitioning some artificial and real life numeric data sets. In the current study we have proposed some SOM based sophisticated genetic operators and then applied those in multiobjective differential evolution based framework to develop some clustering techniques. The newly designed operators are more powerful than existing operators [30] as variable length mating pool sizes are utilized here.

1.3 Motivation

The motivations behind developing the new automatic clustering technique are enumerated below.

a) In recent years SOM [20] has been successfully used in developing some effective reproduction operators for multi-objective evolutionary algorithms like in SOMEA/D [35], SMEA [36] etc. SOM [20] is an unsupervised artificial neural network based model. It maps higher dimensional input vectors to lower dimensional output vectors and thus helps in dimensionality reduction and visualization. Main characteristic of SOM is the preservation of topographical map, that the data samples which are closer in the input space are assigned to the same/nearest neurons in the output space.

In order to use the power of SOM to design some reproduction operators, SOM is first trained using the current population to discover the localities of chromosomes, and then a mating pool is constructed for each solution using the neighborhood relationship extracted by SOM. Thereafter, solutions present in the mating pool are combined using reproduction operation to generate some new solutions.

b) Main limitation of the approach presented in [30] was the use of fixed mating pool size (fixed number of neighboring solutions) for each chromosome in the population. These neighboring solutions for a given solution are identified using SOM [20]. But here, it is adaptive in nature. Different neighborhood sizes are selected based on some probability values. The probability values are determined adaptively based on the performances of generated new solutions over the last few generations. Finally the selected neighborhoods used for construction of mating pool are further used during reproduction operation of DE to generate new solutions.

c) Most of the existing evolutionary algorithms make use of dominance and non-dominance relationships between solutions during the selection operation like non-dominating sorting and crowding distance in NSGA-II [13]. But in this paper, decomposition based selection approach is utilized. The concept of decomposition [35] is used to divide the MOO problem into a number of
single objective subproblems. The objective functions corresponding to these subproblems are optimized in a collaborative manner by the use of MOO. The final solution is determined after considering the objective functions of the subproblems [35].

d) Current paper attempts to develop a multi-objective clustering approach which is a fusion of DE and SOM based operators. Therefore, to better explore the search space and to achieve faster convergence, concept of polynomial mutation [12] is incorporated in the DE based clustering framework to generate more diverse solutions.

Effectiveness of the proposed approach is shown for automatically partitioning five artificial and four real-life data sets. In order to show the efficacy of the proposed clustering technique in handling large data sets, two large scale datasets having more than 5000 data points are also utilized. The proposed multi-objective clustering algorithm also generates a set of non-dominated solutions (Pareto optimal solutions) at the end of final execution. Finally, best solution is selected from this set using Adjusted Rand Index (ARI) [33] which is an external cluster validation measure given that class labels are already known. The obtained partitioning results are compared with several existing clustering techniques like MOCK [15], SMEA_clust [30], MEA_clust [30] which are multi-objective clustering techniques, single objective genetic clustering algorithm (SOGA) [4] and K-means [17]. We have also carried out experiments with out utilizing SOM concept in the proposed framework to illustrate the potentiality of SOM based genetic operators. Results clearly show that the use of neighborhood ensemble extracted from SOM in designing reproduction operator helps in exploring the search space in an effective way. As a real-world application, we choose to apply the proposed clustering technique for scientific article/web-document clustering. Here scientific documents/web documents are clustered based on their semantic similarity. In order to capture meaningful information from documents and also to measure the similarity between two documents, firstly documents are represented as numeric vectors by utilizing the concept of semantic representation, Glove [25]. Thereafter the proposed clustering technique is applied on these numeric vectors to group them into different clusters.

1.4 Key contributions:

Key contributions of the proposed framework are:

1. The developed multi-objective automatic clustering technique utilizes the recently developed SOMEA/D [35] as the underlying optimization strategy. A newly developed SOM based reproduction operator is utilized in the proposed approach which is different from the normal reproduction operator of DE. Some changes are incorporated into DE algorithm to detect clusters automatically.

2. Ensemble of multiple neuron neighborhood sizes is used for construction of mating pool and probabilities of selecting sizes of different neighborhoods are updated based on their performances in generating new improved (good solution in objective space) solutions in the last few generations.

3. Two different internal cluster validity indices named as, Silhouette index [29] and PBM index [24], measuring the goodness of a partitioning are optimized simultaneously using the search capability of multi-objective DE. This simultaneous optimization helps in finding the appropriate number of clusters as well as the optimal partitioning from any given data set.

4. A decomposition based selection operator is utilized in the developed MOO based clustering technique.

5. As an application of the proposed approach, clustering of scientific articles and web documents based on their semantics similarities is illustrated.

2 Self organizing map

Self Organizing Map (SOM) [20] is a type of neural network and can be used as a cluster analysis tool for high dimensional data. It consists of two layers: input and output. Output layer is generally a 2-D grid of neurons. Each neuron is associated with two things: a) position vector in 2-D grid. b) weight vector \( W = w^1, w^2, \ldots, w^D \), where \( D \) is the number of neurons). Main principle behind SOM is that input patterns which are similar in input space appear closer to nearby neurons in the output space. To achieve this, training of SOM is required using input vectors.

Before training of SOM, there is a need to assign a weight vector to each neuron, randomly chosen from the available training data. At each iteration, when an input pattern is presented to the grid, then weight vectors of the winning neuron (closer to presented input pattern) and neighboring neurons are updated to make them closer to the input pattern. In our work, the sequential learning algorithm [16, 31] is utilized for the training of SOM as shown in Algorithm 1. This algorithm returns the updated weight vectors of different map units at the output. Here, training data for SOM will be the solutions present in the population which are discussed in next Section 3.2.
In this paper, we have proposed an efficient automated decomposition based multi-objective clustering technique, named as SOMDEA.clust, which is a hybridization of self-organizing map and differential evolution algorithm and is shown in Fig 1. The notations used and basic operations of SOMDEA.clust are described below.

3.1 Mathematical notations used in the algorithm:

- \( N \): Population size
- \( C_h_i \): \( i \)th chromosome or solution
- \( \mathcal{N} \): Number of data points
- \( d \): Dimension of a data point
- \( g_{\text{max}} \): Maximum number of generations
- \( V = \{v^1, v^2, \ldots, v^\mathcal{X}\} \): Ensemble of different neuron neighborhood sizes (NNS)
- \( \beta^i,k \): Probability of generating solution using \( k \)th NNS at \( t \)th generation
- \( D, \sigma_0, \eta_0 \): Number of neurons, initial neighborhood size and initial learning rate, respectively.
- \( \mathcal{X} \): randomly \( \mathcal{X} \) indices generated, \( idx^k(k = 1, \ldots, \mathcal{X}) \)
- \( W = \{w^1, w^2, \ldots, w^D\} \), weight vector of \( D \) neurons
- \( G \): History length
- \( f_i(y) \): \( i \)th objective function value of solution \( y \)
- \( Q \): Mating pool
- \( \text{CR} \) and \( \text{MP} \): Crossover probability and mutation probability used in DE algorithm
- \( K_i \): Number of clusters exist for \( i \)th solution \( C_h_i \)

3.2 Solution representation and population initialization

First step of the proposed approach is to initialize the population containing a large number of solutions. To initialize a solution, procedure used in the paper [30] is deployed. A solution ‘i’ (or chromosome denoted by \( C_h_i \)) is represented in the form of tuple as \( < \text{cluster centers}, K_i, PBM, SI, L > \), where \( K_i \) is the number of clusters, \( \text{cluster centers} \) comprises of a set of real numbers which represent the coordinates of possible cluster centers, \( L \) is the set of labels predicted after application of any clustering algorithm, PBM and SI are the objective functional values. As the proposed approach is used to find the appropriate number of clusters, therefore for each solution \( C_h_i \), \( K_i \) is varied between 2 to \( \sqrt{\mathcal{X}} \), where \( i = 1, 2 \cdots N \) and \( N \) is the total number of solutions.

Thus, population consists of fix number of chromosomes (say \( N \)) with varying number of clusters and is represented as \( P = \{C_h_1, C_h_2, \ldots, C_h_N\} \).
The variable length chromosomes also participate in SOM training to generate the mating pool; and weight vector of neuron and input vector should be of same dimension. In order to make this feasible, solutions are converted into fixed length vectors by padding some zeros at the end. Series of steps to create an initial solution are shown in Fig. 2.

**Example:** Let \( N = 9, d = 2 \) and for chromosome \( \mathcal{C}h_i \), \( K_i = 2 \), then maximum length of a solution can be \( (\sqrt{9} \times 2 = 6) \). Suppose two cluster centers present in the chromosome are \((2.3, 2.5)\) and \((5.6, 5.7)\). Therefore, the chromosome cluster centers will be viewed as \([2.3, 2.5, 5.6, 5.7, 0, 0]\), where zeros are padded at the end to convert into fixed length vectors. As there are two clusters, therefore each data point can either belongs to cluster 0 or 1. Let predicted labels be \([0, 0, 1, 1, 1, 0, 0, 1, 1]\) and objective functional values of PBM and SI be 0.45 and 0.32, respectively. Then a chromosome \((\mathcal{C}h_i)\) in the form of a tuple looks like

\[
<\{2.3, 2.5, 5.6, 5.7, 0, 0\}, 2, 0.45, 0.32, [0, 0, 1, 1, 1, 0, 0, 1, 1]>
\]

### 3.3 Objective functions used

To optimize the cluster quality, two internal cluster validity indices namely, Silhouette index (SI) [29] and PBM (Pakhira-Bandyopadhyay-Maulik) index [24] are optimized simultaneously using the search capability of DE algorithm. These are same objective functions as used in the paper [30]. Both the objective functions measure the separation and compactness of clusters in different ways and both must be maximized. These objective functions are evaluated after obtaining the partitioning using the cluster centers present in the solution, following the steps discussed in Section 3.2.

The superiority of PBM index over other cluster validity indices, namely, Davies-Bouldin index [9], Dunn index [23] and Xie-Beni index [27] in determining the appropriateness of a partitioning and in dealing with crisp and fuzzy clusters is established in ref. [24]. In ref. [1], Silhouette index is compared with 29 other cluster validity measures (excluding PBM index) namely Davies-Bouldin index [9], Gamma index, C index, Dunn index [23], Xie-Beni index [27] etc. and it was shown to achieve high success rate than others. Inspired by these existing literatures, PBM Index and Silhouette index are incorporated in our proposed framework as the objective functions. Detailed description about these indices are given below:

#### 3.3.1 PBM index:

PBM (Pakhira-Bandyopadhyay-Maulik) is an internal cluster validity index [24] which measures the goodness of partitioning in terms of compactness and cluster separation and should be maximum for a good quality cluster. It can be expressed as

\[
PBM = \left( \frac{1}{K} \sum_{k=1}^{K} E_k \right)^2
\]

Where, \( K \) is the number of clusters, \( N \) is the number of data points, \([\mu_{kj}]_{K \times N}\) is a partition matrix of the data, and \( c_k \) is the center of the \( k \)th cluster, \( E_1 \) represents total scatter of all data points belonging to one cluster, \( D_K \) measures the maximum separation between clusters.

#### 3.3.2 Silhouette index (SI):

SI [29, 32] is an internal cluster validity index used to detect some highly tight/compact clusters which are well-separated from each other. For good quality cluster, it should me maximized. Let \( z_{m1} \) represent the average distance of a point \( x_m \) in \( k \)th cluster to the remaining points of the same
3.4 Selection of neuron neighborhood size

Given the current solution, to form a mating pool, neighborhood size \( v^{pos} \) is required which is selected by ensembling different neuron neighborhood sizes (NNS). Let \( V = \{v^1, ... , v^X\} \) which is an ensemble of different neuron neighborhood sizes. The approach for determining NNS for current solution having index ‘q’ is as follows:

- At each generation randomly \( X \) indices, \( idx^k \) \((k = 1, ... , X) \) are generated in the range of \([1, N]\), where \( N \) is the number of solutions in the population.
- If \( q = idx^k \), \( k \in [1, ... , X] \), then \( v^{pos} = v^k \). Otherwise, \( v^{pos} \) from \( V \) having highest probability \( \in \{\beta_1^1, \beta_2^1, ... , \beta_1^X\} \) is selected. The probability generation depends on several other factors like probability of generating improved (better than current) solution using \( k^{th} \) NNS over last few generations and it is updated at each generation.

Thus, each solution can select different NNS. In general, if current solution index lies in randomly selected \( X \) indices then corresponding NNS is selected from \( V \). Otherwise, out of different NNS, only that NNS is selected which generates good solutions over last few generations. This implies exploration of solutions in good region which helps in convergence towards optimal solutions.

**Example:** Let \( N = 12 \), \( X = 4 \), \( \beta = (0.25, 0.25, 0.20, 0.30) \), \( V = (4, 6, 8, 10) \).

**Case (a)** If current solution index ‘q’ exists in idx: Let \( idx = (2, 1, 3, 5) \) and we want to generate new solution for first (q = 1) solution. As \( q = 1 \) exists in idx at index 2, therefore corresponding neuron neighborhood size \( v^{pos} \) will be \( v^2 \), i.e., 6.

**Case (b)** If current solution index ‘q’ does not exist in idx: Let \( idx = (2, 8, 3, 5) \). As \( q = 1 \) does not exist in idx and the maximum probability in \( \beta \) can be found at index 4, therefore corresponding neighborhood size \( v^{pos} \) will be \( v^4 \), i.e., 10.

3.5 Mating pool construction

Mating pool (Q) includes those solutions which can mate to generate new solution using genetic operators like crossover etc. For its construction, detailed algorithm is presented in Algorithm 2 [35]. Given the current solution \( \mathcal{C} h_q \), first its closest neuron \( b \) in the trained SOM is identified using shortest distance criteria (Line 2), i.e., the Euclidean distance between the current solution and weight vectors (\( W \)) of different neurons on SOM are computed as shown in (5). The neuron having lesser distance will be selected as closest neuron, \( b \).

\[
b = \arg \min_{1 \leq u \leq D} \| \mathcal{C} h_q - w^u \|_2
\]

Here in (5), \( D \) is the number of neurons, \( u \) is the index of neuron, \( w^u \) is the weight of \( u^{th} \) neuron. Thereafter \( v^{pos} \) neighboring neurons of \( b \) are selected, where \( v^{pos} \) is the size of neighborhood (number of neighboring neurons) (Line 3–4). Finally, all solutions associated with neighboring neurons are collected to form a mating pool for current solution (Line 5–6). As principle of SOM states that similar solutions will be mapped to nearby neurons, this property helps in collecting similar solutions in the mating pool. In the proposed framework as we are using different neuron neighborhood sizes, therefore each solution can have different sizes of mating pool.

In Algorithm 2, \( P \): the population containing solutions \( \{\mathcal{C} h_1, \mathcal{C} h_2, ... , \mathcal{C} h_N\} \), \( q \) is the current solution number, \( DIST \): distance matrix formed using position vectors of neurons in the grid, \( v^{pos} \): neuron neighborhood size, \( L \): list of winning neuron indices for different solutions in population \( P \), \( \mathcal{C} h_q \): current solution for which the mating pool is generated, \( Q \) be the mating pool.

**Algorithm 2 MatingPool_Construction(q, DIST, P, \mathcal{C} h_q, L, v^{pos})**

1. \( Q = \emptyset \)
2. Find winning neuron ‘b’ for current solution \( \mathcal{C} h_q \) based on minimum Euclidean distance according to (5).
3. sort \( b^{th} \) row of \( DIST \) in ascending order and record the sorted indices in \( J \).
4. Find out \( v^{pos} \) neighboring neurons: \( B = (J(2) \cdots J(v^{pos} + 1)) \)
5. for \( i = 1 \) to \( v^{pos} \) do
6. \( Q = Q \cup \{\mathcal{C} h_q\} \cup L(e) = B(i), e = 1, 2 \cdots N \)
end
7. \( \text{return } Q \) for solution \( \mathcal{C} h_q \)
3.6 Solution generation

The details of new solution generation are given in Algorithm 3. Here, new solution generation means generation of new cluster centers. The procedure generates new solution, $y$, mainly using the current solution, $\mathcal{C}h_q$, and its mating pool, Q, generated using Algorithm 2. It first generates a trial solution using DE crossover operator [26] (Line 2) and then repairing takes place to make the trial solution feasible using upper and lower boundaries of the population (Line 3). Next, polynomial mutation operator [12] is applied on repaired trial solution to make the solution more diverse (Line 4) followed by again repairing (Line 5). At last, any of the three types of mutation will be applied (Line 6) to explore the search space as our proposed algorithm deals with finding the number of clusters from the dataset in an automatic manner.

These types of mutations are: normal mutation, insertion mutation and deletion mutation. Any mutation is selected based on some probability called as mutation probability (MP) and chosen based on literature survey [27].

Let $\eta_m$ and $p_m$ be polynomial mutation and distribution index, respectively. Then algorithm for new solution generation is shown Algorithm 3 [35]. It should be noted that during any computation for new solution generation, only real feature values (excluding padding zeros) of the current solution ($\mathcal{C}hq$) are considered.

Example: Let $K_q = 3, d = 2$, current solution $\mathcal{C}h_q = \{s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, 0, 0\}$, and Q (Mating Pool) consist of three solutions which are $\{s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, 0, 0\}$, $\{s_{41}, s_{42}, s_{43}, s_{44}, 0, 0, 0, 0\}$, $\{s_{31}, s_{32}, s_{33}, s_{34}, s_{35}, s_{36}, s_{37}, s_{38}\}$. Then at the time of generating a trial solution $y' \text{(Step-2)}$ only $K_q \times d = 3 \times 2 = 6$ features of all the solutions are considered as the current solution has only 6 features. The remaining features are treated as zero as shown in Fig. 3.

Final step of Algorithm 3 where solution undergoes normal, insertion or delete mutation based on some random probability, MP, is shown in Fig. 4.

It should be noted that in case of (a) normal mutation, number of clusters for new solution will remain same as $K_q$. (b) insertion mutation, number of clusters for new solution increases by 1, i.e., $\{K_q + 1\}$. (c) delete mutation, number of clusters for new solution decreases by 1, i.e., $\{K_q - 1\}$. After generating new solution, following additional steps are performed to obtain the final solution in tuple form.

1. New solution generated using Algorithm 3 passes into K-means as initial cluster centers with corresponding no. of clusters (say $K_{new}$).
2. Clusters center and predicted labels obtained after executing K-means algorithm, treated as cluster centers and L as a part of tuple. Also, in the tuple, number of clusters, K will be $K_{new}$.

3. Now, PBM and SI index values are calculated to obtain the complete solution in tuple form.

### 3.7 Updating the probability of selecting NNS

As discussed in Section 3.4, if current solution index $q \neq idx^k$, then an appropriate neighborhood size (NNS) is selected out of ensemble of different neuron neighborhood sizes depending on the probability of generating good solution over last few generations (history length) and updated at each generation. Let $\beta_{t,k}$ be the probability of generating solution using $k$th NNS at generation $t$, $\rho_{t,k}$ indicate the number of solutions that can be improved by the new solution, $cnt_{t,k}$ be the number of new solutions generated using $v^k$ where $v^k$ is the $k$th neuron neighborhood size and history length $= G$. Then method of updating probability is given below.

1. Calculate the utility of $v^k$, $k = 1 \ldots K$

   $$u_{t,k} = \begin{cases} \frac{\sum_{j=1}^{t} \rho_{t,j}}{\sum_{j=1}^{t} cnt_{t,j}} & \text{if } t < G \\ \frac{\sum_{j=t-G+1}^{t} \rho_{t,j}}{\sum_{j=t-G+1}^{t} cnt_{t,j}} & \text{Otherwise} \end{cases}$$

2. Calculation of the probability for $k = 1 \ldots K$ at $t + 1$ generation

   $$\beta_{t+1,k} = \frac{u_{t,k} + \epsilon}{\sum_{j=1}^{K} u_{t,j} + \epsilon}$$

**Fig. 4** Generation of new solution. Here $\text{rand}()$ is a function which generates some random number between 0 to 1.
3.8 Update solution

To update the existing solutions with the newly generated y and to achieve faster convergence, decomposition based selection [35] scheme is employed. Detailed algorithm to update the solution is shown in Algorithm 4. In this approach, Tchebycheff decomposition [21] is used to build N single objective optimization problems (Line 2–5). Before updating any solution in archive A, y is first used to update \( z^* \) (Line 1). Then for \( p^{th} \) subproblem, \( p = 1, 2, \ldots N \), the function values \( g(x^p) \), \( g(y) \) and improvable degree, \( idg(p) = g(y) - g(x^p) \) are calculated (Line 2–6).

Finally, \( idg \) is sorted in descending order and the solution \( x \) with largest improvement is replaced by solution \( y \) (Line 7–8). In Algorithm 1, \( q \) is the current solution number, \( A \) is the archive containing copy of the population, \( \lambda \) = \( \lambda_1^p \lambda_2^p \ldots \lambda_m^p \) is a set of weight vectors, \( \sum_{j=1}^{m} \lambda_j^p = 1 \), \( z^* = (z_1^*, z_2^*, \ldots z_m^*) \) is the reference point, \( m \) is the number of objective functions considered. In our problem, \( z^* \) is initialized by \( z_j^* = \text{max}(f_j(x)|x \in P_0), j = 1, 2 \ldots m \), where, \( P_0 \) is the initial population.

Algorithm 4 Update Solution(A, y, N, z*, q)
1: Update ideal point \( z^* \): If \( z_{j1}^* \neq f_j(y), j=1,2,\ldots m \)
2: for \( p=1 \) to \( N \), \( \lambda^p = (\lambda_1^p \lambda_2^p \ldots \lambda_m^p) \in \Delta, x^p \in A \) do
3: Calculate the objective value of subproblem \( p \) for \( x^p \) and \( y \) as below:
4: \( g(y|\lambda^p, z^*) = \text{max} \{x^p y \leq m \lambda_j^p \} \times |f_j(y) - z_j^*| \)
5: \( g(x^p|\lambda^p, z^*) = \text{max} \{x^p y \leq m \lambda_j^p \} \times |f_j(x^p) - z_j^*| \)
6: Calculate of improvable degree: \( idg = g(y|\lambda^p, z^*) - g(x^p|\lambda^p, z^*) \)
7: Sort \( idg \) in descending order and store sorted indices in \( J \).
8: if \( idg(J(k)) > 0 \), then Replace \( x^{J(k)} \in A; x^{J(k)} = y \)
9: return \( z^* \), A

Let \( \omega^p = (\omega_1^p, \omega_2^p \ldots \omega_m^p) \), be a set of uniformly distributed vectors such that \( \omega_j^p > 0 \) and \( \sum_{j=1}^{m} \omega_j^p = 1 \) for \( p = 1, 2 \ldots N \). In actual decomposition based approach [37], the weights \( \omega^p \) are directly used to construct the subproblems. But, recent studies [14] say that subproblems constructed using \( \lambda^p \) can give rise to better distributed final population, where \( \lambda^p \) can be expressed mathematically as below:

\[
\lambda_j^p = \frac{1}{\omega_j^p} - \frac{1}{\sum_{k=1}^{m} \omega_k^p}
\]

where, \( p = 1, 2, \ldots N, j = 1, 2, \ldots m \). Thus, Tchebycheff decomposition [21] with \( \lambda^p \) is adopted in our framework.

3.9 Terminating condition

The termination condition of any multi-objective algorithm could be 1) number of iterations until no improvement on best solution is reached or 2) maximum number of objective function evaluations, \( g_{\max} \), is reached. In our proposed framework, last one is used as the termination criterion.

3.10 Selection of best solution

All multi-objective algorithms generate a set of non-dominated (equally important) solutions on the final Pareto optimal front. But sometimes, decision maker wants to select one solution from the available set of solutions. Therefore, in the current paper a single best solution is selected based on external validity measure, named as Adjusted Rand Index (ARI) [33] given that class labels are known for a given data set. It makes use of true partitioning and obtained partitioning and then calculates similarity between these. ARI definition states that higher the value it has, better is the partitioning. It’s range is [-1,1]. Thus ARI is computed for all the solutions (or partitioning solutions) available on the final Pareto Optimal Front and then solution having maximum ARI is reported here.

4 Datasets used

Experimental results of the proposed clustering approach are shown for four real-life data sets and five artificial data sets. Artificial data sets include, \( AD_{5.2} \), \( Square_4 \), \( Sizes_5 \), \( Spherical_{4.3} \) and \( Spherical_{6.2} \). Real-life data sets are \( Iris \), \( BreastCancer \), \( LiverDisorder \) and \( Newthyroid \). All real-life data sets are publicly available in UCI repository [22]. Number of actual clusters and total number of data points present in these data sets are provided in Table 1. Actual partitioning diagrams before application of any clustering algorithm on artificial data sets namely (a) \( AD_{5.2} \), (b) \( Square_4 \), (c) \( Sizes_5 \), (d) \( Spherical_{4.3} \), (e) \( Spherical_{6.2} \), are shown in Fig. 5.
Table 1 Results on different data sets by proposed approach (PA1): SOMDEA$_{clust}$, SMEA$_{clust}$, PA2: MDEA$_{clust}$, MEA$_{clust}$, MOCK, SOGA, K-means clustering technique

<table>
<thead>
<tr>
<th>Data set</th>
<th>$N$</th>
<th>$d$</th>
<th>AC</th>
<th>PA1</th>
<th>SMEA$_{clust}$</th>
<th>PA2</th>
<th>MEA$_{clust}$</th>
<th>MOCK</th>
<th>SOGA</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OC</td>
<td>MS</td>
<td>OC</td>
<td>MS</td>
<td>OC</td>
<td>MS</td>
<td>OC</td>
</tr>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0.60</td>
<td>2</td>
<td>0.72</td>
<td>2</td>
<td>0.82</td>
<td>3</td>
</tr>
<tr>
<td>Cancer(Breast)</td>
<td>699</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>0.33</td>
<td>5</td>
<td>0.76</td>
<td>2</td>
<td>0.37</td>
<td>2</td>
</tr>
<tr>
<td>Newthyroid</td>
<td>215</td>
<td>5</td>
<td>3</td>
<td>0.44</td>
<td></td>
<td>5</td>
<td>0.61</td>
<td>4</td>
<td>0.59</td>
<td>2</td>
</tr>
<tr>
<td>LiveDisorder</td>
<td>345</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>0.98</td>
<td>8</td>
<td>0.99</td>
<td>3</td>
<td>0.98</td>
<td>2</td>
</tr>
<tr>
<td>AD$_{5,2}$</td>
<td>250</td>
<td>2</td>
<td>5</td>
<td>0.25</td>
<td>5</td>
<td>5</td>
<td>0.35</td>
<td>6</td>
<td>0.39</td>
<td>5</td>
</tr>
<tr>
<td>Square4</td>
<td>1000</td>
<td>2</td>
<td>4</td>
<td>0.56</td>
<td>4</td>
<td>4</td>
<td>0.56</td>
<td>4</td>
<td>0.56</td>
<td>4</td>
</tr>
<tr>
<td>Sizes5</td>
<td>300</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Spherical$_{6,2}$</td>
<td>400</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>0.17</td>
<td>6</td>
</tr>
</tbody>
</table>

Here, $N$: Number of data points, $d$: number of features (dimension) in input data, AC: actual number of clusters in the data set.

5 Experimental setup and discussion of results

This section reports about the comparing methods, parameter setting, and experimental results followed by statistical significance of results obtained.

5.1 Comparing methods

For the purpose of comparison, our proposed clustering framework is compared with various state-of-the-art clustering techniques. The approaches selected for comparison are MOO based clustering approach namely, MDEA$_{clust}$, SMEA$_{clust}$, MEA$_{clust}$, MOCK, SOGA which is a Single Objective Genetic Algorithm based clustering [4] and traditional clustering technique, K-means [17]. K-means clustering algorithm is simple and well-known clustering algorithm having limited computational complexity. To run K-means, the number of clusters present should be known beforehand. Note that our proposed clustering technique is automatic in nature. It determines the number of clusters from a given data set in an automatic way. In order to make a fair comparison, for K-means, the number of clusters is fixed to $K$ where $K$ is the value of the optimal number of clusters determined by the proposed approach, SOMDEA$_{clust}$.

In order to quantify the obtained clustering results by different comparing algorithms, an external cluster validity index, Minkowski Score [33], is used. It is an external cluster validity measure used to validate the clustering results. Lower the MS score, better is the partitioning. Its minimum optimal value is '0'. In our approach, MS is reported for that specific solution which have maximum ARI out of multiple solutions.

The results reported in this section are the average values over 20 runs. All the approaches were implemented on a Intel Core i7 CPU 3.60 GHz with 4 GB of RAM on Ubuntu in python environment.

5.1.1 SMEA$_{clust}$

SMEA$_{clust}$ [30] is a multi-objective clustering technique developed using the fusion of SOM and multi-objective differential evolution framework. It optimizes two objective functions, PBM and SI, simultaneously and is able to detect the optimal number of clusters as well as the optimal partitioning from a data set in an automatic way. It suffers from the problem of fixed mating pool size.

5.1.2 MEA$_{clust}$

MEA$_{clust}$ [30] approach is similar to SMEA$_{clust}$, but it does not utilize any SOM based operators.

5.1.3 MDEA$_{clust}$

Another version of our proposed multi-objective clustering, namely, MDEA$_{clust}$, without utilizing the SOM based genetic operators is developed in this paper. It is also able...
Fig. 5 Actual partitioning of artificial data sets a AD_{5,2}, b Square4, c Sizes5, d Spherical_{4,3}, e Spherical_{6,2}

Fig. 6 Actual partitioning of a cluto-t7-10k, b cluto-t8-8k
Obtained partitionings after application of proposed clustering technique on artificial data sets a $AD_{5,2}$, b $Square4$, c $Sizes5$, d $Spherical_{4,3}$, e $Spherical_{6,2}$

Table 2 Results on data sets, cluto-t7-10k and cluto-t8-8k, by the proposed approach (PA1): SOMDEA$_{clust}$, SMEA$_{clust}$, PA2: MDEA$_{clust}$, MEA$_{clust}$ and, K-means clustering technique

<table>
<thead>
<tr>
<th>Data set</th>
<th>$N$</th>
<th>d</th>
<th>AC</th>
<th>PA1</th>
<th>SMEA$_{clust}$</th>
<th>PA2</th>
<th>MEA$_{clust}$</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OC</td>
<td>MS</td>
<td>OC</td>
<td>MS</td>
<td>OC</td>
</tr>
<tr>
<td>cluto-t7-10k</td>
<td>9208</td>
<td>2</td>
<td>9</td>
<td>5</td>
<td>0.9229</td>
<td>5</td>
<td>0.9229</td>
<td>6</td>
</tr>
<tr>
<td>cluto-t8-8k</td>
<td>7677</td>
<td>2</td>
<td>8</td>
<td>14</td>
<td>0.8879</td>
<td>5</td>
<td>0.9828</td>
<td>5</td>
</tr>
</tbody>
</table>

Here, $N$: Number of data points, d: number of features (dimension) in input data, AC: actual number of clusters in the data set
to detect the appropriate number of clusters automatically from a given data set and optimizes PBM-index [24] and Silhouette index [29] simultaneously. It is developed to show the effectiveness of our newly designed genetic operators utilizing SOM based neighbor information.

5.1.4 MOCK

MOCK [15] is a multi-objective clustering algorithm with automatic K-determination and optimizes two objective functions, connectedness and compactness simultaneously to improve the cluster quality.

5.1.5 SOGA

SOGA [4] is a single objective clustering technique where single cluster validity index is optimized using the search capabilities of genetic algorithm.

In our experiments, SOGA based clustering was executed multiple times with the number of clusters varying between 2 to $\sqrt{N}$, where $N$ is the number of data points. The final partitioning is selected based on the minimum value of MS score.

5.1.6 K-means

K-means [17] is a well known unsupervised clustering algorithm. It assumes that the number of clusters (K) is known a priori. Here, the given dataset is partitioned into K clusters by using the procedure of minimum center distance based criterion. A specific point is assigned to that cluster with respect to which it is having the minimum distance.

5.2 Experiment Settings

Different parameters used for conducting experiment with algorithm SOMDEA,clust are as follows. These parameters are selected after conducting a detailed sensitivity analysis.

- Population size (N): 50
- SOM parameter: initial learning rate ($\eta_0$) = 0.8, initial neighborhood size ($\sigma_0$) = 0.1, Number of neurons (D) = $|N|/2$. Number of iterations for training = $|N|$.
- DE control parameter: F = 0.5 and CR = 0.8
- $N = 5$ and ensemble of neuron neighborhood size $\{u_1, u_2, u_3, u_4, u_5\}$ = $\{4, 5, 6, 8, 10\}$
- Initial probabilities of selecting NNS, $\{\beta^{0.1}, \beta^{0.2}, \beta^{0.3}, \beta^{0.4}, \beta^{0.5}\}$ = $\{0.25, 0.25, 0.25, 0.25, 0.25\}$
- Mutation probability (MP): normal = [0, 0.6], insertion = [0.6, 0.8], deletion = [0.8, 1.0]
- Polynomial mutation distribution index ($\eta_m$) = 21 and probability ($p_m$) = 0.6

Same parameters are used in MDEA,clust approach (excluding SOM parameters). For K-means, number of clusters is taken as obtained by the proposed approach. SOGA [4] and MOCK [15] are executed with default parameters (codes of the algorithms are provided by the authors).

5.3 Results of artificial data sets:

It is evident from the results reported in Table 1 that the proposed clustering algorithm, SOMDEA,clust, is able to find the appropriate number of clusters as well as the optimal partitioning from all the data sets used here. For data sets Spherical_6.2 and Spherical_4.3, the five automatic multi-objective clustering algorithms, SOMDEA,clust, MDEA,clust, SMEA,clust, MEA,clust and MOCK are able to identify the optimal partitioning as both these data sets contain some well-separated clusters; the results are also comparable to those obtained by K-means clustering. But single objective genetic clustering algorithm fails for these data sets. For data sets, Sizes5 and AD_5.2, MOCK fails to detect the appropriate number of clusters. But, SOMDEA,clust and SMEA,clust perform similarly for this data set and better than others in terms of number of clusters detected and MS score.

For Square4, SOMDEA,clust, SMEA,clust, MEA,clust and MDEA,clust perform similarly but SOGA performs poorly. Final partitionings obtained after application of SOMDEA,clust for AD_10.2, AD_5.2, Square4, Sizes5, Spherical_4.3 and Spherical_6.2 data sets are shown in Fig. 7.

Results corresponding to datasets, cluto-t7-10k and cluto-t8-8k are shown in Table 2. Here, for cluto-t8-8k dataset, SOMDEA,clust performs better than others. However, the number of clusters obtained by SOMDEA,clust is more than actual as this dataset is having clusters of varying densities and arbitrary shapes. While for cluto-t7-10k dataset, MDEA,clust performs better than others. Here also, it’s not able to detect the actual number of clusters because of same reason as in cluto-t7-10k. But, in general, the techniques SOMDEA,clust and MDEA,clust, shows that ensemble of different neighborhood sizes for designing genetic operations for the multi-objective clustering technique helps in improving the results. The obtained partitionings corresponding to the best results obtained for cluto-t7-10k and cluto-t8-8k, are shown in Fig. 8. Note that we have excluded MOCK from comparison because it is not capable of handling large sized chromosomes,
i.e., as the number of data points increases, length of the solution/chromosome increases. Also, it can detect only well-separated cluster [15]. SOGA is also not able to handle such large datasets. Pareto Optimal solutions obtained by our proposed approach (SOMDEA_{clust}) on cluto-t7-10k and cluto-t8-8k datasets are shown in Fig. 11a and b, respectively. Note that based on domain knowledge, the number of clusters is varied in the range of 2 to 30. Population size and the maximum number of generations are set as 30 and 20, respectively.

**Results of real-life data sets:** For Iris data set, SOMDEA_{clust}, MDEA_{clust}, SMEA_{clust}, MEA_{clust} and SOGA all are able to detect the appropriate number of clusters. But, MS scores attained by SOMDEA_{clust} and MDEA_{clust} are equal and those are comparable to the value obtained by K-means. For Breast Cancer and Newthyroid data sets, our approach performs better than others in terms of Minkowski score.

For LiveDisorder data set, proposed approach and SMEA_{clust} both have low MS scores in comparison to other approaches. In general, both SOM based approaches, SOMDEA_{clust} and SMEA_{clust} are able to detect the appropriate number of clusters, which clearly show the utility of employing SOM based operators in multi-objective optimization framework. But, due to use of ensemble of neighborhood sizes, our approach achieves the lowest MS scores for Iris, Breast Cancer and Newthyroid data sets in comparison to other multi-objective clustering techniques. Due to high dimensionality of real-life data sets, visualization of partitioning results is not shown.

### 5.4 Student’s t-test

To check the superiority of proposed clustering algorithm, statistical hypothesis test named as Welch’s t-test is conducted, guided by [34] at the 5% significance level. It checks whether the improvements obtained by the

<table>
<thead>
<tr>
<th>Data set</th>
<th>SMEA_{clust}</th>
<th>MDEA_{clust}</th>
<th>MEA_{clust}</th>
<th>MOCK</th>
<th>SOGA</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>1.99E-089</td>
<td>0.788494</td>
<td>7.10E-196</td>
<td>0</td>
<td>0</td>
<td>0.788494</td>
</tr>
<tr>
<td>Cancer (Breast)</td>
<td>0.000677</td>
<td>1.19E-069</td>
<td>5.30E-035</td>
<td>2.74E-021</td>
<td>9.16E-011</td>
<td>0.000677</td>
</tr>
<tr>
<td>Newthyroid</td>
<td>7.10E-196</td>
<td>2.01E-296</td>
<td>1.14E-257</td>
<td>0</td>
<td>3.65E-110</td>
<td>3.65E-110</td>
</tr>
<tr>
<td>LiveDisorder</td>
<td>0.788494</td>
<td>0.000677</td>
<td>0.788494</td>
<td>0.788494</td>
<td>3.65E-110</td>
<td>3.65E-110</td>
</tr>
<tr>
<td>AD_{5,2}</td>
<td>0.788494</td>
<td>5.35E-153</td>
<td>5.34E-153</td>
<td>1.61E-237</td>
<td>0</td>
<td>3.65E-110</td>
</tr>
<tr>
<td>Square4</td>
<td>0.788494</td>
<td>0.788494</td>
<td>0.788494</td>
<td>5.30E-35</td>
<td>0</td>
<td>3.65E-110</td>
</tr>
<tr>
<td>Sizes5</td>
<td>0.788494</td>
<td>0.000677</td>
<td>2.74E-21</td>
<td>0</td>
<td>0</td>
<td>0.788494</td>
</tr>
<tr>
<td>Spherical_{6,2}</td>
<td>0.788494</td>
<td>0.788494</td>
<td>0.788494</td>
<td>0.788494</td>
<td>0</td>
<td>0.788494</td>
</tr>
<tr>
<td>Spherical_{4,3}</td>
<td>0.788494</td>
<td>0.788494</td>
<td>0.788494</td>
<td>0.788494</td>
<td>2.01E-296</td>
<td>0.788494</td>
</tr>
</tbody>
</table>
proposed approach have happened by chance or those are statistically significant. t-test provides p-value. Smallest p-value indicates that the proposed approach is better than others. In our experiment, p-values are calculated considering two groups. One group contains the list of Minkowski score (MS) values generated by proposed algorithm. While another group contains the list of MS values generated by some other algorithm. In this t-test, two hypotheses are considered: null hypothesis and alternative hypothesis. First hypothesis is that there is no significance difference between median values of two groups. On the other hand, alternative hypothesis shows that there is significant change between median values of two groups. We have shown the p-values obtained in Table 3 which clearly support the results of Table 1.

6 Application of proposed approach in text clustering

To show the real-life application of the proposed decomposition based multi-objective automatic clustering, we have chosen the problem of text document clustering. Two text data sets are used which are explained below in brief:

1. NIPS 2015: The scientific articles published in Neural Information Processing Systems (NIPS) conference which is an important core ranked conference in the machine learning domain are collected. Data set is named as ‘NIPS 2015’ (downloaded from kaggle.com site) and contains 403 articles (written in English) of the year 2015. Word Cloud of NIPS 2015 articles is shown in Fig. 9 which is generated using paper text and is sufficient to capture topics related to published articles. It includes research articles in the field of deep learning, computer vision, cognitive science, reinforcement learning etc. Among many attributes given in the data set, only title, abstract and paper text are used in our experiments. Here, topmost 8, 25 and 150 words are selected from title, abstract and paper text, respectively.

2. WebKB: In WebKB [7] dataset, text documents are web pages, not scientific articles. It consists of web pages collected from computer science department of 4 different universities, which are Texas, Cornell, Wisconsin and Washington. In this paper, we have used total of 2803 documents out of 4199 documents. The corresponding Word cloud is shown in Fig. 10.

6.1 Representation scheme used

To represent a text document (for NIPS 2015 and WebKB datasets) in the form of a numeric vector, semantic representation namely, Glove [25] is used with different dimensions. Glove learns word representation by constructing a co-occurrence matrix that basically counts how frequently a word appears in a context and then this matrix is reduced to lower dimension, where, each row represents a word vector. For our experiment, we have used the pre-trained Golve word representations of

---

Table 4 Obtained clustering results by different algorithms

<table>
<thead>
<tr>
<th>Data set</th>
<th>#DP</th>
<th>R</th>
<th>#d</th>
<th>PA1</th>
<th>SMEA_clust</th>
<th>PA2</th>
<th>MEA_clust</th>
<th>MOCK</th>
<th>SOGA</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIPS 2015</td>
<td>403</td>
<td>Glove</td>
<td>100</td>
<td>17</td>
<td><strong>0.4289</strong></td>
<td>0.3884</td>
<td>0.4278</td>
<td>0.3714</td>
<td>0.3464</td>
<td>0.3833</td>
</tr>
<tr>
<td></td>
<td>WebKB</td>
<td>2803</td>
<td>Glove</td>
<td>300</td>
<td>12.148</td>
<td><strong>0.6325</strong></td>
<td>0.1037</td>
<td>20.017</td>
<td>2.00</td>
<td>39.00</td>
</tr>
</tbody>
</table>

(a) Table reporting Dunn index values

<table>
<thead>
<tr>
<th>Data set</th>
<th>#DP</th>
<th>R</th>
<th>#d</th>
<th>PA1</th>
<th>SMEA_clust</th>
<th>PA2</th>
<th>MEA_clust</th>
<th>MOCK</th>
<th>SOGA</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS 2015</td>
<td>403</td>
<td>Glove</td>
<td>100</td>
<td>17</td>
<td><strong>0.4289</strong></td>
<td>0.3884</td>
<td>0.4278</td>
<td>0.3714</td>
<td>0.3464</td>
<td>0.3833</td>
</tr>
<tr>
<td>WebKB</td>
<td>2803</td>
<td>Glove</td>
<td>300</td>
<td>12.148</td>
<td><strong>0.6325</strong></td>
<td>0.1037</td>
<td>20.017</td>
<td>2.00</td>
<td>39.00</td>
<td>4.106</td>
</tr>
</tbody>
</table>

(b) Table reporting values of numerator (m) and denominator (n) of Dunn index values reported in Table 4a. Here, m: minimum distance between two points belonging to different clusters, n: maximum diameter amongst different clusters

(c) Table reporting DB index values

<table>
<thead>
<tr>
<th>Data set</th>
<th>#DP</th>
<th>R</th>
<th>#d</th>
<th>PA1</th>
<th>SMEA_clust</th>
<th>PA2</th>
<th>MEA_clust</th>
<th>MOCK</th>
<th>SOGA</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS 2015</td>
<td>403</td>
<td>Glove</td>
<td>100</td>
<td>17</td>
<td><strong>1.4631</strong></td>
<td>1.5774</td>
<td>1.6239</td>
<td>1.6561</td>
<td>1.7448</td>
<td>2.1561</td>
</tr>
<tr>
<td>WebKB</td>
<td>2803</td>
<td>Glove</td>
<td>300</td>
<td>12.148</td>
<td><strong>1.3501</strong></td>
<td>1.3932</td>
<td>2.4429</td>
<td>1.3938</td>
<td>2.0660</td>
<td>2.1561</td>
</tr>
</tbody>
</table>

Here, PA1 and PA2 are SOMDEA_clust and MDEA_clust, respectively, #DP: number of scientific articles, R: representation used, d: number of features, OC: obtained clusters, DI: Dunn index, DB: Davies-Bouldin index

50, 100, 200 and 300 dimensions available at Github.2

To generate the document vector, vectors (obtained from Glove representation) of words present in the document are averaged.

6.2 Evaluation measure:

For the evaluation of results, we have used two internal cluster validity indices, namely, Dunn Index (DI) [23] and Davies-Bouldin (DB) index [10]. Both measures the compactness and separation of clusters, but, in a different way.

1. Dunn index: Let \( \mathcal{S} \) be any clustering algorithm. Then it can be expressed mathematically as

\[
DI(\mathcal{S}) = \frac{\min_{C_k, C_l \in \mathcal{S}, C_k \neq C_l} (\min_{i \in C_k, j \in C_l} dist(i, j))}{\max_{C_m \in \mathcal{S}} diam(C_m)}
\]

Where, numerator denotes the shortest distance between two points belonging to different clusters, denominator represents the largest intra-cluster distance, \( diam(C_m) \) is the diameter of \( m \)th cluster. The higher value of Dunn index indicates better clustering.

2. DB index: It is defined as follows:

\[
DB = \frac{1}{K} \sum_{i=1}^{K} D_i
\]

Where,

\[
D_i = \max_{i \neq j} R_{i,j} \quad \text{and} \quad R_{i,j} = \frac{S_i + S_j}{M_{i,j}}
\]

Where, \( M_{i,j} \) is the separation between the \( i \)th cluster and the \( j \)th cluster, \( S_i \): within-cluster scatter for cluster \( i \); \( K \): number of clusters.

6.3 Experiments results and discussions

In Table 4, the best results obtained by different comparing methods are reported after varying the Glove dimension. Table 4a reports the Dunn index values. As can be seen from Table 4a, the best value of Dunn index obtained by our approach, SOMDEA_clust, for NIPS2015 dataset is 0.4289. While for, WebKB dataset, it is 0.6325 which is obtained by SMEA_clust and MEA_clust approaches. As our proposed algorithm is the improved version of SMEA_clust, we have tried to find out the reason behind their good performance for WebKB dataset. Therefore, we also reported the numerator and denominator values corresponding to obtained Dunn index values in Table 4b. Here, numerator denotes the minimum distance between two points belonging to different clusters, while denominator denotes the maximum diameter amongst different clusters. Based on the analysis, it was found that Dunn index is biased towards clusters having more number of data points which in turn, decreases the cluster compactness and increases the cluster separations. Out of 4
clustering technique on a cluto-t7-10k, b cluto-t8-8k, c NIPS 2015 data set using glove with 100 dimensions and d WebKB data set using Glove 300 dimensions. Here, $x$ – axis denotes values of Silhouette index and $y$ – axis denotes values of PBM Index.

7 Conclusions and future works

In this paper a new self-organized decomposition based multi-objective clustering technique, SOMDEA_clust, is developed. The proposed approach is a fusion of differential evolution and SOM. Here, SOM, trained using solutions in the current population, is used to find out neighborhood information, which is further utilized for generating the mating pool. This mating pool further takes part in genetic operations. The use of SOM for genetic operations and ensemble of different neighborhood sizes helps in exploring the search space efficiently. Two objective functions namely, PBM and silhouette index are optimized simultaneously in order to obtain the optimal partitioning. Concept of polynomial mutation is used to generate the highly diverse solution in DE framework. Proposed framework is applied on seven artificial data sets and four real-life data sets. In
a part of the paper the effectiveness of incorporating SOM based reproduction operator is also established. Another technique, namely MDEA_clust, is also developed in this paper without making use of SOM based reproduction operator to show the effectiveness of SOM in our proposed approach. As a real-life application, proposed multi-objective clustering technique is also applied on clustering of scientific articles as well as on a set of web documents. In future, we would like to apply this technique to solve some other real-life problems like search result clustering, multi-document text-summarization etc. Effect of different sets of objectives in the proposed framework can be another future research direction. Automatic adaption of various parameters used in the proposed approach can also be another future research direction.

Acknowledgments This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Affiliations

Naveen Saini¹ · Sriparna Saha¹ · Aditya Harsh² · Pushpak Bhattacharyya¹

¹ Department of Computer Science and Engineering, Indian Institute of Technology, Patna, Bihar, India
² UPES, Dehradun, India
AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES:

Q1. Authors biography and photo are desired. Please provide. Otherwise, please confirm if unnecessary.
Q2. Naveen Saini has been set as the corresponding author. Please check and advise if correct.
Q3. Please provide significance for bold entries found in Tables 1, 2 and 4. Otherwise, please remove emphasis.
Q4. Please check table 4 if captured/presented correctly.
Q5. Please provide volume number for ref. [10].