Vol.67: e24230508, 2024
https://doi.org/10.1590/1678-4324-2024230508
ISSN 1678-4324 Online Edition

BABT
Brazilian Archives of Biology and Technology

Article - Engineering, Technology and Techniques

Advancing Gene Expression Data Analysis: an Innovative Multi-objective Optimization Algorithm for Simultaneous Feature Selection and Clustering

Pooja Gupta[1*]
https://orcid.org/0000-0001-7285-5516

Abhay Kumar Alok[2]
https://orcid.org/0000-0002-0940-4330

Vineet Sharma[3]
https://orcid.org/0000-0003-4737-4272

[1]Dr. A. P. J. Abdul Kalam Technical University, Scholar, Lucknow, Uttar Pradesh, India; [2]Indian Institute of Technology, Patna, India; [3]KIET Group of Institutions, Delhi-NCR, Ghaziabad, India.

Editor-in-Chief: Alexandre Rasi Aoki
Associate Editor: Raja Soosaimarian Peter Raj

Received: 18-Jun-2023; Accepted: 29-Sep-2023

*Correspondence: pooja.02.gupta@gmail.com; Tel.: +91-8860672035 (P. G.)

HIGHLIGHTS

- A novel multi-objective optimization algorithm is proposed for simultaneous feature selection and gene data clustering.
- An efficient feature selection approach is utilized for relevant feature subset in gene data clustering.
- An approach based on simulated annealing is employed to simultaneously optimize three objective functions.
- The obtained Clustering results are proven superior to nine other existing clustering techniques.

Abstract: Clustering algorithms play a crucial role in identifying co-expressed genes in microarray data, while feature subset identification is equally important when dealing with large data matrices. In this research paper, we address the problem of simultaneous feature selection and gene expression data clustering within a multi-objective optimization framework. Our approach employs the Archived multi-objective simulated annealing (AMOSA) algorithm to optimize a multi-objective function that incorporates two internal validity indices and a feature weight index. To determine data point membership in different clusters, we utilize a point symmetry-based distance metric. We demonstrate the effectiveness of our proposed approach on three publicly available gene expression datasets using the Silhouette index. Furthermore, we compare the clustering results of our approach, unsupervised feature selection and clustering using Multi-objective optimization framework (UFSC-MOO), to nine other existing techniques, showing its superior performance. Statistical significance is confirmed through Wilcoxon Rank Sum test. Also, biological significance test is employed to show that the obtained clustering solutions are biologically enriched.

Keywords: Gene expression data Clustering; Feature selection; Point symmetry based distance; AMOSA; Cluster validity index; Feature weight index.

INTRODUCTION

Classification and Clustering are the major areas of machine learning that captivate researchers for in-depth study and application. The classification techniques are widely applied for text classification [1-6], sentiment analysis [7-9], sarcasm identification [10, 11], evaluation of reviews by multiple instructors [12], and many more. The clustering technique groups the data based on similarity in implicit property of data. However, Classification and clustering become a challenging task for high dimensional data because of high computational complexity. To this end, multiple meta-heuristic feature selection schemes are proposed for classification. Xue Yu and coauthors [13] addresses the problem of feature selection in large scale feature set for classification. For this, they designed a self-adaptive particle swarm optimization (SaPSO) algorithm for feature selection with multiple candidate solution generation strategy (CSGS) that outperforms various competing techniques in terms of classification accuracy. Besides, Song X.F. and coauthors [14] also addresses the aforementioned challenges of high dimensional data set, thereby designed an algorithm based on correlation-guided clustering and particle swarm optimization showing competing accuracy. Inconsistent data holding missing values or noisy data in large data set may lead to local optimum solutions. To this end, particle swarm optimization based feature selection using fuzzy clustering [15] is employed for class imbalance problem. Besides, consensus clustering can be utilized to handle imbalanced class data by applying undersampling approach to majority class [16]. Deep learning approaches like Recursive Neural Network [17] and Evolutionary algorithm like genetic algorithm [18] are also employed for extracting relevant features for sentiment classification. Here, high dimensional data such as gene expression dataset is typically represented as a large matrix, where each row corresponds to gene expression levels and each column represents different experimental conditions or samples. Thereby**,** Microarray data clustering poses a significant challenge due to its high dimensionality, high computational complexity and the need to capture the behaviour of thousands of genes simultaneously. Additionally, appropriate sample selection can aid in visually representing gene behaviour across various experimental conditions. So, feature selection is a crucial step in microarray data clustering to achieve precise and meaningful visual representations of gene interactions. Mostly feature selection approaches employ wrapper method that may hold certain limitations- the requirement of class information, commonly based on single validity measure, unable to retain diverse solutions and hence, cannot avoid local optima [19]. With this view, several feature selection methods [20] for gene expression data analysis are pointed by researchers. To this end, the top ranked features are selected by using- the pipelining framework of attribute clustering and feature ranking techniques [21]; the hybrid approach based on the ReliefF filter method and a novel meta-heuristic Equilibrium Optimizer (EO) [22]; the threshold based sample selection decided using the division operation of relational algebra [23]. While clustering approaches applied to reduced sample spaces have shown promising results still, they may fall short when dealing with symmetric, overlapping, and high-dimensional data, particularly when optimizing a single objective function [24, 25]. Also, these approaches may find difficulty in avoiding local optimal solutions. To overcome this issue, various multi-objective approaches have been proposed for simultaneous feature selection and clustering. Prakash J. and coauthors [26] proposed a multi-objective based gravitational search and K-means algorithms for simultaneous feature selection and clustering. Here, K-means is employed for centroid initialisation; though this approach has shown encouraging results, it faces high computational cost. Hancer E. and coauthors [19] contributed by using a multi-objective differential evolution technique based on variable-string length encoding scheme and have shown promising results for real world data set. More recent research work for simultaneous feature selection and gene data clustering employing the different techniques includes- fuzzy clustering using non-dominated sorting genetic algorithm II and multi-objective evolutionary algorithm based on decomposition [27]; Co-expressed genes identification using Archived Multi-objective Simulated Annealing [28]; transcriptome-wide time series expression profiling [29]; multi-view clustering using ensemble technique [30]; a-Priori Biological Knowledge in clustering [31]. Further, Focussing on the need to deal with noisy constraints in constraint set of gene data, Wang Z. and coauthors [32] proposed the multi-objective clustering algorithm using semi-supervised learning, based on selection of constraints and multi-source constraints integration. Apart from clustering, several meta-heuristic approaches are embedded for feature selection and classification in bio-medical domain as well; utilization of cuckoo search for feature subset selection to increase the accuracy of Naive Bayes classifier [33]; artificial ant colony optimization with naive bayes classifier for classification of cancer while, cuckoo search is applied for feature subset selection [34]; A hybrid approach using cuckoo search (CS) for minimising the number of elected genes and increasing the performance of Naive bayes classifier [35].

In the previous studies, clustering methods applied to reduced sample spaces have demonstrated potential. However, they may fail short when dealing with complex data scenarios like symmetry, overlap, and high dimensionality, particularly when optimizing a single objective functions. In light of these observations, multi-objective clustering approaches have emerged, aimed at obtaining optimal clusters from gene expression data. However, there remains a gap in the literature regarding the simultaneous resolution of feature selection and clustering challenges within a multi-objective optimization framework. Motivated by this observation, we propose an approach that formulates the problem of sample selection and clustering as a multi-objective optimization problem with preserving the biological significance of gene clusters. To optimize multiple objective functions, we employ a simulated annealing-based multi-objective optimization method called AMOSA [36]. This approach is a response to the limitation of conventional methods that overlook certain data characteristics.

Our proposed approach UFSC-MOO, tackles the task of multi-objective clustering and feature selection. Here, we encode cluster centers and features as intermediate solutions and use AMOSA to simultaneously optimize the Sym index [37], XB index [38], and feature weight index. This optimization process aims to identify the Pareto-optimal front [36], which represents a set of non-dominating solutions. These solutions correspond to different cluster centres and combinations of samples with respect to genes, ultimately leading to well-separated gene clusters. To obtain symmetrical and compact clustering solutions, the assignment of data points to different clusters is determined based on point symmetry based distance measure [39]. To evaluate the performance of UFSC-MOO, we conduct experiments on three publicly available gene expression datasets. Further, we compare the clustering results obtained using UFSC-MOO with nine other popular existing clustering approaches, including FCM [40], MO-fuzzy [41], MOGA [42], SGA [43], SOM [44], Hierarchical average linkage clustering [45], CRC [46], K-mean [47], and Spectral clustering [48]. The performance of UFSC-MOO is assessed using the silhouette index. Additionally, we employ wilcoxon rank sum test [49] and biological significance test [50] to demonstrate the statistical significance and biological significance of the gene clustering results obtained through our proposed technique, UFSC_MOO. In the end of this research, the contribution of all authors is equally important. P.G. designed the study, analysed and interpreted the data, implemented the algorithm, validated the results statistically, compared the results to other existing techniques, drafted the manuscript and; A.A. contributed in supervision of experiments, result validation, paper proofing, concept validation; and V.S. made contributions in critical revision of the manuscript, supervision of experiments, providing administrative, technical and material support.

MATERIAL AND METHOD

In the present section, we discuss the proposed multi-objective approach (UFSC-MOO) for simultaneous feature selection and gene expression data clustering. The experimental set up for simulation of proposed methodology is given as- the data pre-processing steps, the computation of fitness function (Sym index, XB index and feature weight index), implementation of the AMOSA underlying multi-objective optimization technique, and various parameter settings are integrated within the C environment. Statistical significance tests and all the visualizations, such as the Eisen plot and cluster profile plot for gene expression data analysis are conducted using the Matlab environment. Biological significance tests are performed using the Gene ontology tool [51].

Problem Statement

Let the gene data matrix is given by X = {$\overline{X}_j$: j=1, 2….n} where, $\overline{X}_j$ is represented as a D-dimensional vector. Here the goal is to assign data points to K distinct clusters by determining the membership value Z_{kj} , which represents the degree of membership of jth point to kth cluster given as, $\sum_{k=1}^{K} \sum_{j=1}^{n} Z_{kj}$ = n. Secondly, including the entire given feature set in clustering may lead to curse of dimensionality. So, projecting the D-dimensional feature space into F-dimensional feature space while, gratifying the various cluster quality measures is posed as the multi-objective optimization problem here. Further, the proposed work can be mathematically formulated as below:

Input: The input is taken as a set of n number of data points given by X= {$\overline{X}_j$: j=1,2….n}, where $\overline{X}_j$ is represented as a D-dimensional vector.

Optimization framework: The multiple cluster validity measures are optimised simultaneously according to search strategy of AMOSA [36], a multi-objective optimization approach.

Output: The objective is to output an efficient feature subset F such that, F <=D and clustering is performed on the selected F features for the given gene data set to be partitioned into K different clusters. The membership matrix Z with dimension K X n is generated dynamically, that represents the membership value of data points to the particular cluster. Here, K is the number of clusters and n is the total number of data points. The membership value is given by,

$$\sum_{k=1}^{K} \sum_{j=1}^{n} Z_{kj} = n, \tag{1}$$

Where, If $Z_{kj} = 0$, then jth point is not a member of cluster k
If $Z_{kj} = 1$, then jth point is member of cluster k

Proposed Approach: UFSC-MOO - Simultaneous Unsupervised Feature Selection and Clustering in a Multi-Objective Optimization framework

In the present section, we have discussed the step-by-step working of proposed algorithm, UFSC-MOO. Here, three objective functions are optimised simultaneously utilizing AMOSA [36]. Consequently, we obtain a solution set comprising several dominating and non-dominating solutions. These non-dominating solution set form a Pareto-optimal front. Figure 1 shows the stepwise procedure of UFSC-MOO, followed by description of each steps in detail.

Pre-processing of Gene expression data

The experimental analysis utilizes three benchmark gene expression datasets as shown in Table 1 - Yeast Sporulation [52], Yeast cell cycle [53], and Arabidopsis Thaliana [54]. The details of these data sets and their pre-processing are discussed as below:

Yeasts Sporulation: Initially, there are total 6118 genes and their expression values are measured over 7 time points. After pre-processing step, total 474 active genes are selected. The pre-processing step includes log-transformation and calculating the root mean square values, using a threshold of 1.60.

Yeast Cell Cycle: With 6000 genes measured over 17 time points, 384 active genes are selected after pre-processing step. Similar to pre-processing of Yeasts Sporulation, log-transformation and root mean square values are computed, applying a threshold of 1.60.

Arabidopsis Thaliana: Initially, there are total 138 pre-processed genes measured over 8 time points. In pre-processing step, we have eliminated entire rows having zero attributes. After this step, we normalized all the data values with mean zero and variance one.

These pre-processing steps are specific to each dataset and aim to prepare the gene expression data for further analysis.

Table 1. Description of pre-processed data set where N and D represent count on genes and samples respectively

Data Set	**N**	**D**
Arabidopsis Thaliana	138	8
Yeast Sporulation	474	7
Yeast cell cycle	384	17

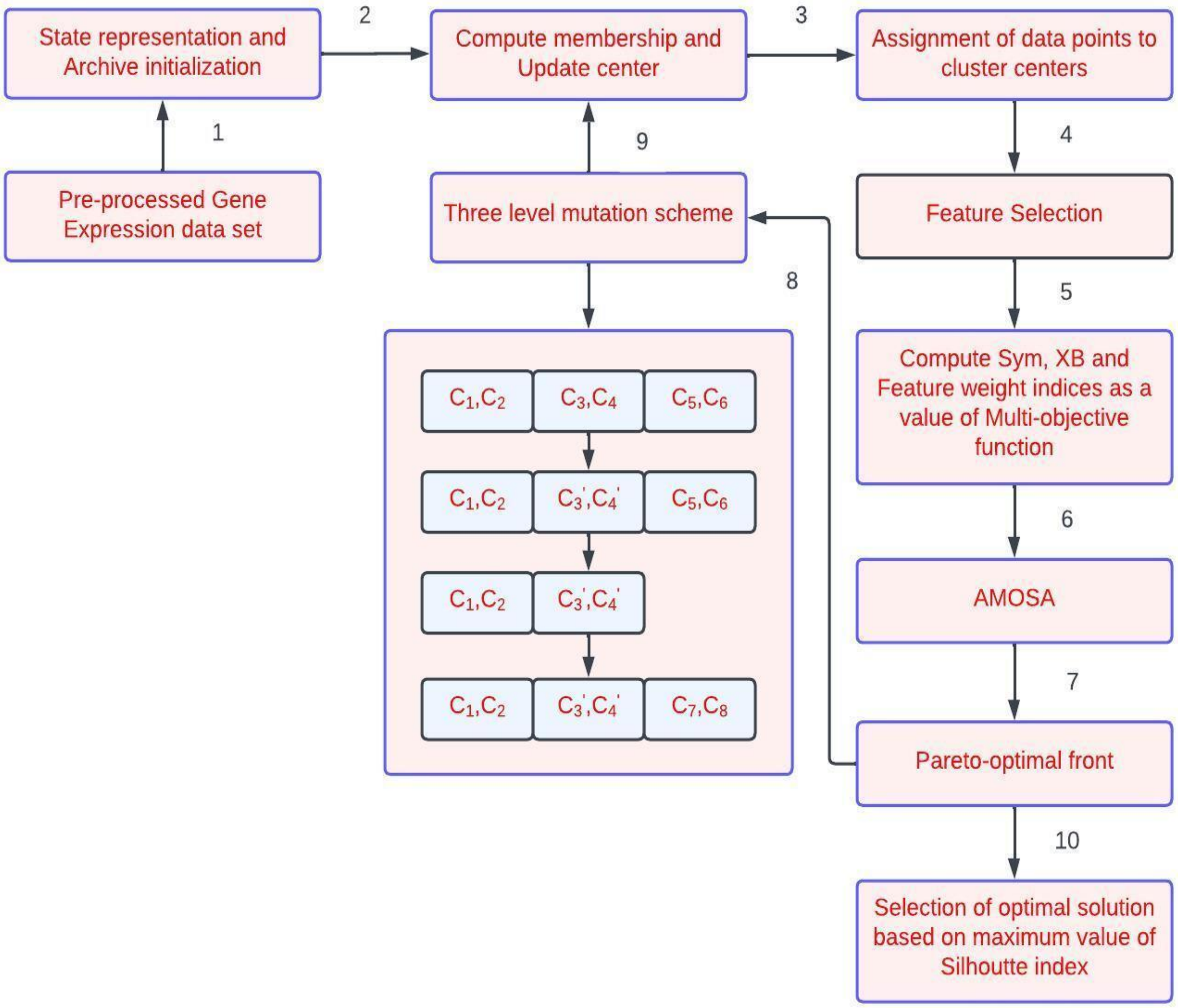

State representation and Archive initialization
2
Compute membership and Update center
3
Assignment of data points to cluster centers
1
9
4
Pre-processed Gene Expression data set
Three level mutation scheme
Feature Selection
8
5
C_1,C_2
C_3,C_4
C_5,C_6
Compute Sym, XB and Feature weight indices as a value of Multi-objective function
C_1,C_2
C_3',C_4'
C_5,C_6
6
C_1,C_2
C_3',C_4'
AMOSA
7
C_1,C_2
C_3',C_4'
C_7,C_8
Pareto-optimal front
10
Selection of optimal solution based on maximum value of Silhoutte index

Figure 1. Working principle of UFSC-MOO

Encoding scheme and archive initialization

Here, the state representation of AMOSA [36] basically comprises two fields. The first field is encoded as a set of real numbers showing coordinates of cluster centres. While, second one is given by string of decimal values between 0 and 1 exhibiting the feature activation code (feature weight) of samples in different combinations. The larger the value of activation code, the more preferred is corresponding feature, that is 0 indicates that the feature is ignored while 1 indicates that the feature is preferred. The length of encoded string can be determined by (F+K) x F, where F and K represent the number of features and the number of clusters respectively. Figure 2 shows an illustration for state representation, comprising three clusters (K) and five different features (F). Here, we take threshold value as 0.5, thereby selecting the features F1, F3, F4 having activation code (feature weight) value greater than pre-decided threshold value. Now, data instance allocation to clusters and calculation of objective functions is done considering only the selected features.

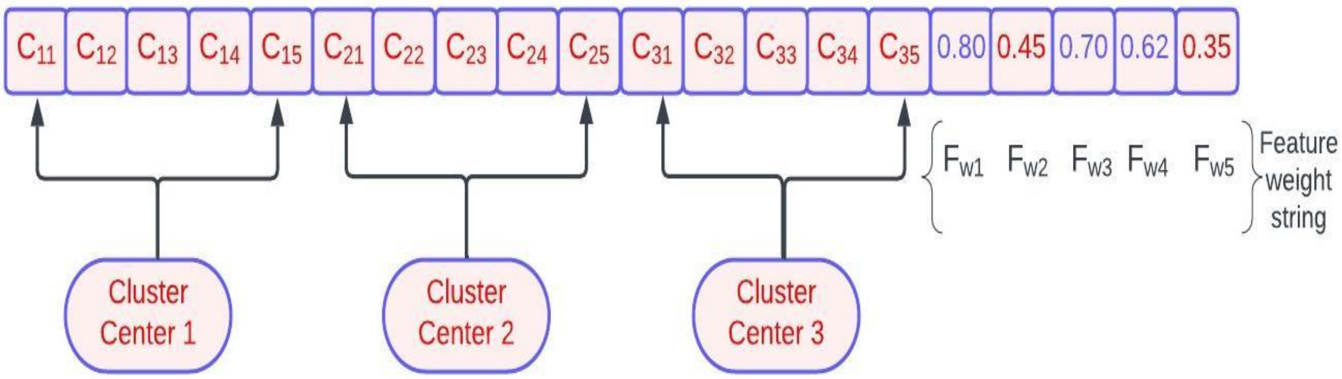


Figure 2. An Illustrative example for state representation

Assignment of points and membership updating

Let the cluster centres represent the different clusters. In UFSC-MOO, the allocation of data points to the various clusters is determined based on point symmetry based distance measure, D_{p_s} [39]. The advantage of utilizing point symmetry based distance metric [39] is that we obtain more symmetrical and compact clustering solutions. In a particular encoded string, let K be the total number clusters, F represented as number of features, n is the total number of genes and cluster centre is given by C. The assignment of data points to cluster centres is done as follows:

$$\bar{X}_i \ \varepsilon \text{ cluster } k, \ \forall \ 1 \le i \le n,$$

$$\text{If } D_{p\_s}(\bar{X}_i, \bar{C}_k) \le D_{p\_s}(\bar{X}_i, \bar{C}_j), \ \forall \ 1 \le j \le K, \ j \ne k,$$

$$\text{And } D_{sym}(\bar{X}_i, \bar{C}_k) = D_{p\_s}(\bar{X}_i, \bar{C}_k) \ / \ D_{euc}(\bar{X}_i, \bar{C}_k) \le \alpha, \tag{2}$$

In second case, if $D_{p\_s}(\bar{X}_i, \bar{C}_k) / D_{euc}(\bar{X}_i, \bar{C}_k) > \alpha$, then K-means algorithm is utilized to assign point $\bar{X}_i$ to any cluster z. The value of z is decided using the following condition:

$$D_{euc}(\bar{X}_i, \bar{C}_z) \le D_{euc}(\bar{X}_i, \bar{C}_j), \text{ where } 1 \le j \le K, \ j \ne z \tag{3}$$

Here, the symmetry of data point with respect to cluster centre within a cluster is calculated by D_{sym}. If value of D_{sym} is small, then the point is considered almost symmetrical to cluster centre. For this purpose, we have taken α as a threshold to examine the symmetricity property. If the amount of symmetricity is less than α, then we can conclude that the given data point is symmetrical, otherwise the data point is not considered symmetrical to any cluster. In that case, we employ K-means for assignment of points to cluster according to minimum value of Euclidean distance metric given by D_{euc}. Here, UFSC-MOO computes the distance for only those features that holds larger value of feature weight when compared to threshold value in a particular string. To accomplish this, the parameter α is set as the maximum nearest neighbour distance among all the points within a given data set.

Feature Selection

Feature selection is responsible for selecting the relevant features, thereby decreasing computational complexity in high dimensional gene expression data set. Here, features are assigned normalized activation code (feature weight) ranging between 0 and 1 based on feature importance. Here, 0 signifies feature is omitted, while 1 signifies feature is selected. We have considered a threshold activation code value as 0.5, that shows if the value of feature activation code is greater than 0.5, the feature is included in selected feature subset, otherwise the feature is discarded. Further, the value of absolute feature weight (F_{wt}) is calculated as below and is optimised in multi-objective function by AMOSA [36] in order to obtain optimal clusters.

$$F_{wt} = \frac{D-d}{D-1} \tag{4}$$

Where D is the total number of features and d is the number of selected features based on value of their activation codes. So, feature selection is carried out in each iteration as the part of multi-objective optimisation by AMOSA, unless the desired clustering solutions are obtained.

Multi-objective optimization framework and the objective functions used

The proposed approach UFSC-MOO, employs a simulated annealing based multi-objective optimization technique, called AMOSA [36]. This algorithm encompasses the idea of archive consisting the set of equally valuable solutions. The addition of new solution to archive or deletion/ replacement of existing solutions from archive are based on dominance relation of new solution with current solution according to some set of criteria [36]. The amount of dominance between two solutions, say x and y is represented as:

$$\Delta \mathrm{Dom}_{x,y} = \prod_{\mathrm{i=1,fi(x)\neq fi(y)}}^{\mathrm{N}} \frac{|\mathrm{fi(x)-fi(y)}|}{\mathrm{Ri}}, \tag{5}$$

Where, N represents total number of objective functions and R_i corresponds to the range of i^{th} objective function. These parameters are essential for computing the acceptance probability of new solution to the archive. The archive in this approach is constrained to contain well-distributed pareto-optimal solutions. It is subject to two limits, namely soft limit and hard limit. The Soft limit (SL) decides the upper threshold on the number of non-dominated solutions to be kept in archive during the process. If the number of generated non-dominated solutions exceeds SL, then archive size is reduced to hard limit (HL) by applying clustering method. The HL serves as a strict maximum size for archive at the end of algorithm. The experiment for the proposed UFSC-MOO clustering technique acquires several parameter settings including SL=100, HL=50 and num_iter = 50, Tmax=100, Tmin=0.00001, cooling rate α = 0.9, probability of mutation=0.2 and probability of crossover =0.8. Now, selection of the best solution from multiple competing solutions in archive is based on the value of an internal cluster validity index. Further, optimization of the three objective functions is performed simultaneously using AMOSA, which includes the two internal cluster validity indices and the feature weight index. The details of AMOSA algorithm [36] and objective functions are given as:

AMOSA Algorithm

```
Set Tmax, Tmin, HL, SL, no_iter, α, Temp = Tmax
Initialize the Archive.
curr_point = random(Archive). /* a solution is randomly chosen from the Archive*/
while (Temp > Tmin)
    for (i=0; i< no_iter; i++)
        new_point=perturb(curr_point).
        Check the domination status of new-point and curr_point.
        /* dominance code for three different cases */
        If (current-pt dominates new-pt) /* Case 1*/
            $\Delta\mathrm{dom}_{avg} = \frac{(\sum_{i=1}^{K} \Delta\mathrm{dom}_{\mathrm{i,new\_point}}) + \Delta\mathrm{dom}_{\mathrm{curr\_point,new\_point}}}{K+1}$
            /* K = total number of points in the Archive which dominate new_point, K ≥ 0*/
            Prob = $\frac{1}{1+\exp(\Delta\mathrm{dom}_{\mathrm{avg}} * \mathrm{Temp})}$.
            Set new-point as current-point with probability=Prob
        If (curr_point and new_point are equally non-dominating) /* Case 2*/
            Check the domination status of new_point and points in the Archive.
            if (new_point is dominated by K points in the Archive, where (K ≥1))/*Case 2(i)*/
                Prob = $\frac{1}{1+\exp(\Delta\mathrm{dom}_{\mathrm{avg}} * \mathrm{Temp})}$.
                $\Delta\mathrm{dom}_{avg} = \frac{(\sum_{i=1}^{K} \Delta\mathrm{dom}_{\mathrm{i,new\_point}})}{\mathrm{K}}$
                Set new_point as current_point with probability=Prob.
```

Cont.

```
                if (new_point is non-dominating to all other points in the Archive) /*Case 2(ii)*/
                    Set new_point as current_point and append new_point to the Archive.
                    If size of Archive > SL
                        Cluster Archive to HL number of clusters/*Reduce the size of
                    archive to HL*/
                If (new_point dominates K points of the Archive, where K ≥1) /* Case 2(iii)*/
                    Set new_point as current_point and append it to the Archive.
                    Remove all the K dominated points from the Archive.
            If (new_point dominates curr_point) /* Case 3 */
                Check the domination status of new_point and points in the Archive.
                If (new_point dominates K points of the Archive, where K ≥1) /* Case 3(i)*/
                    $\Delta dom_{min}$ = minimum of the difference of domination amounts between the
                    new_point and the K points
                    Prob = $\frac{1}{1+\exp(-\Delta \mathrm{dom}_{\mathrm{min}})}$.
                    Set point of the archive which corresponds to $\Delta dom_{min}$ as curr_point with
                    probability = Prob
                    else set new_point as curr_point.
                If (new_point is non-dominating with respect to the points in the Archive)/*Case
                3(ii)*/
                    Set new_point as the curr_point and append it to the Archive.
                    If curr_point is in the Archive, remove it from the Archive.
                    Else if Archive_size> SL.
                        Cluster Archive to HL number of clusters.
                If (new_point dominates K other points in the Archive ) /* Case 3(iii)*/
                    Set new_point as current_point and add it to the Archive.
                    Remove all the K dominated points from the Archive.
        End for Temp= α∗Temp.
End while
If Archive-size > SL
Cluster Archive to HL number of cluster
```

Objective functions

The specific details of the used objective functions are as follows-

Sym Index: Sym index quantifies the average symmetry in relation to cluster centres, providing a measure of overall symmetry [39]. It is based on point symmetry distance. It seeks for highly symmetric clusters. Let the given data set Z= $\{\bar{z}_i : \mathrm{i} = 1{,}2..\mathrm{n}\}$ is partitioned into K well-separated clusters. For each j clusters, cluster centre $\bar{C}_j$ is calculated by: $\bar{C}_j = \frac{\sum_{k=1}^{n_j} \bar{z}_{ij}}{n_j}$. The total compactness within clusters represented by $\mathcal{E}_k$ is defined by: $\mathcal{E}_k = \sum_{I=1}^{K} E_l$. Here, E is the total symmetrical deviation for some cluster j and is given by $E_j = \sum_{i=1}^{n_j} D*_{p\_s}(\bar{z}_{ij}, \bar{C}_j)$. Where, $D_{p\_s}(\bar{z}_{ij}, \bar{C}_j)$ is computed as product of $D_{sym}(\bar{z}_{ij}, \bar{C}_j)$ and $D_{euc}(\bar{z}_{ij}, \bar{C}_j)$. The separation between two cluster centroids given by D_k, should be maximised.

$$D_k = \max || \bar{C}_i - \bar{C}_j || \quad \forall\, i, j \text{ such that, } 1 \le i, j \le K, \tag{6}$$

With a goal to identify highly symmetric and well-separated clusters while keeping a count on number of clusters, the value of Sym index is maximised. Sym Index can be expressed as:

$$\mathrm{Sym}\ (K) = \left(\frac{1}{K}\ \mathrm{X}\ \frac{1}{\varepsilon_k}\mathrm{X}\ D_k\right), \tag{7}$$

XB Index: XB cluster validity index [38] basically focuses on two characteristics of clusters-Compactness and separation. The good partitioning scheme tries to attain lower value of compactness and larger value of separation between cluster centres. XB index can be defined as ratio of compaction to the separation computed in terms of Euclidean distance, derived as:

$$XB= \sum_{a=1}^{K} \sum_{b=1}^{n} \mu_{ab}^{2} || \overline{x}_b - \overline{c}_a ||^2 /(n * (\min_{a \neq k} || \overline{c}_a - \overline{c}_k ||^2)), \tag{8}$$

Where, n and K represent the number of data points and the total number of clusters encoded in a solution respectively. Here, $\overline{c}_a$ denotes a^th cluster, $\overline{x}_b$ denotes b^th data point and μ_{ab} denotes belongingness of the data point b to cluster a. If μ_{ab} holds 1, indicates data point b belongs to cluster a, while value 0 shows that data point b does not belong to cluster a. However, XB index seeks for clusters that are hyper spherical in shape and the value of XB index should be minimised to evolve optimal clustering solutions.

Feature Weight Index (F_{wt}): The third index F_{wt} is associated with feature selection scheme that is simultaneously done with clustering procedure. This index is responsible for selection of relevant feature and is represented as feature string in a cluster solution representation. F_{wt} must be maximised to compensate the bias of previous two objective functions on dimensionality. Usually, dataset has tendency to be distributed into a greater number of clusters with smaller area depending on the number of clusters rather than forming lesser number of bigger sized cluster. However, the clustering results may comprise overlapped clusters whenever the first two objective functions cause high dimensionality reduction. The reason for such an issue is that both Sym and XB indices directly or indirectly depends on Euclidean distance for calculation and so are biased towards lower dimensions that lowers the value to 1 [24] [25]. To balance this bias, feature weight index is maximised. Therefore, we can define the overall multi-objective fitness function that is optimized using the popular approach AMOSA as follows:

$$\text{Overall fitness function} = \text{Max}(\text{Sym}, 1/\text{XB}, F_{wt}) \tag{9}$$

The proposed approach UFSC-MOO maximizes the value of aforementioned fitness function using AMOSA framework to obtain highly symmetrical clusters without overlapping and thus we can attain optimal clustering solutions.

Mutation Scheme

Once the optimal set of solutions is obtained, we apply mutation operators to further explore the search space. In UFSC-MOO, each solution comprises two components-cluster component and the selected feature subset component. To maintain diversity within the solutions, we employ mutation operations on the current solution string to generate new solution string. Specifically, we utilize three distinct types of mutation operations on the cluster component of a given solution string, aiming to introduce diversity into the subsequent generation of solutions. The feature string is randomly updated during each iteration. The specific details of these mutation operations are given as:

1. The first mutation operation utilizes the laplacian distribution for modifying the old value of cluster centroids to new value. Here, laplacian distribution is employed to perturb each cluster centroids in solution string using $p(\varepsilon) \propto e^{-|c-\mu|/\delta}$. Here c is the cluster centroid. μ is used to represent position for perturbation and δ is scaling factor initialised as 1. The perturbation operation is applied to all dimensions in the given context.
2. The Second mutation operation involves reducing the number of clusters in a specific solution string by randomly removing one cluster centroid.
3. The third mutation operation is utilized to increase the cluster numbers in a specific solution string. This is done by randomly selecting a data point from the dataset and adding it to the solution string as a new cluster centroid.

The above mutation operations are performed with equal probability on a solution string. Also, any one of these mutation schemes is performed on solution string if is opted for mutation step.

Selection of a single best solution

We obtain a set of non-dominating solutions in pareto-optimal front [37] after completion of all the above stages of proposed UFSC-MOO. These non-dominating solutions are equally good in some or the other prospective. Though all these solutions are equally important yet, we need to find the single best solution, to meet the user requirements and sometimes from comparison prospective. For this, we have employed Silhouette index [55] to select the best solution among all solutions present in pareto-optimal front. The value of Silhouette index can be defined in terms of compaction and separation as:

$$\mathrm{Sil}(C) = \frac{y-x}{(x,y)} \tag{10}$$

Here, x measures cluster compactness calculated as the average distance of genes within a particular cluster; y measures cluster separation computed as the minimum average distance of a gene with respect to other cluster's gene.

RESULTS AND DISCUSSION

In the present section, the results obtained using UFSC-MOO algorithm is discussed. The performance of proposed approach is evaluated and compared with nine other commonly used clustering approaches. Here, three publically available data sets namely Yeast Sporulation, Yeast cell cycle and Arabidopsis Thaliana, are used for experiment after pre-processing step.

Performance Metrics

Here, Silhouette index [55] is used to evaluate the performance of proposed UFSC-MOO technique. The obtained partitioning results are visually shown using cluster profile plot [56] and Eisen plot [56]. Eisen plot is used to represent gene expression value at some time point in an ordinary way. Before plotting them, it is ensured that genes belonging to same clusters are placed sequentially together. For this, we reorder the data when required. Here, firstly we seek for the colour in data matrix which is exactly similar to its spotted colour in microarray. Different colours show different expression levels. Here, red colour represents higher expression levels. Green colour shows low expression levels and black colour is dominated when no values for differential expression is found while, white colour represents separation between two clusters. Cluster profile plot is generally used for visualizing a mean profile plot for each cluster in a cluster analysis. Before cluster profile plotting, initially the average gene expression values are computed with respect to corresponding time points for each cluster. Then, we compute standard deviation of expression values of different points within a particular cluster. Additionally, we conducted Wilcoxon rank sum [26] test to assess the statistical significance of obtained clustering solutions. To check the biological significance, we referred Gene ontology annotation [51] database.

Result Analysis

The proposed technique UFSC-MOO is primarily built upon the idea of simulated annealing, specifically AMOSA algorithm, to enable simultaneous feature selection and unsupervised clustering. We have applied this technique on three publically available gene expression data sets, utilizing them for both feature selection and gene clustering simultaneously. To evaluate the performance of UFSC-MOO, we compared its results with nine widely used clustering techniques- FCM [40], MO-fuzzy [41], MOGA [42], SGA [43], SOM [44], Hierarchical average linkage clustering [45], CRC [46], K-mean [47] and Spectral clustering [48]. Figure 3 shows the entire steps of experiment. The assignment of genes to different clusters is done based on the distance metrics computed using all available time points. In a similar way, objective function calculation also took into account these time points. Table 2 shows cluster count and the elected time points that are found by UFSC-MOO technique. Additionally, we evaluated the clustering performance by employing Silhouette index value. The higher value of Silhouette index indicates the better partitioning outcomes. To compare UFSC-MOO with other popular clustering techniques, we examined the silhouette index values obtained from each method. Table 3 shows the comparative analysis of UFSC MOO with aforementioned clustering approaches.

Table 2. Partitioning results comprising selected features, count on clusters and Silhouette index value as determined by UFSC-MOO.

Data Set	Preferred Features	K	Sil(C)
Arabidopsis Thaliana	1,3,5,6,8	4	0.4412
Yeast Sporulation	1,3,4, 6,7	6	0.6331
Yeast cell cycle	1,3,5,6,7,8,12,13,16,17	5	0.4621

Table 3. Comparative analysis of UFSC-MOO with other clustering techniques via Silhouette index value

Algorithm	Sporulation		Cell Cycle		Thaliana	
	K	Sil(C)	K	Sil(C)	K	Sil(C)
UFSC_MOO	6	0.6331	5	0.4621	4	0.4412
MO-fuzzy	6	0.5961	5	0.4472	4	0.4162
Multi-objective GA	6	0.5781	5	0.4265	4	0.4055
Fuzzy c-means	6	0.4751	6	0.3685	4	0.3691
Simple GA	6	0.5712	5	0.4251	4	0.3971
Average Linkage	6	0.5045	4	0.4385	5	0.3095
Self-organizing map	6	0.5812	6	0.3971	5	0.2261
CRC	7	0.5665	5	0.4212	4	0.4061
Spectral Clustering	6	0.5595	4	0.3565	4	0.1591
K-mean	6	0.4578	5	0.4085	5	0.3695

Results by proposed method

For Yeast Sporulation gene data, UFSC-MOO clustering approach selects five out of 7 features and based on these features the six clusters are evolved (K = 6). The five samples are selected (given in Table 2). The obtained value of Sil(C) is 0.6331 which is found highest when compared to nine other clustering approaches. The number of clusters and their corresponding S(C) scores for different clustering techniques are as follows: MO-fuzzy (6, 0.5961), MOGA (6, 0.5781), FCM (6, 0.4751), SGA (6, 0.5712), Average Linkage (6, 0.5045), SOM (6, 0.5812), CRC (7, 0.5665), Spectral (6, 0.5595) and K-mean (6, 0.4578). For Yeast Cell Cycle gene data, UFSC-MOO clustering approach selects ten out of seventeen features and based on these features the five clusters are evolved (K = 5). The ten selected samples are as given in Table 2. The obtained value of Sil(C) is 0.4621 which is found highest when compared to other nine clustering approaches. The cluster count and Sil(C) scores for the clustering techniques are obtained as: MO-fuzzy (5, 0.4472), MOGA (5, 0.4265), FCM (6, 0.3685), SGA (5, 0.4251), Average Linkage (4, 0.4385), SOM (6, 0.3971), CRC (5, 0.4212), Spectral (4, 0.3565) and K-mean (5, 0.4085) clustering techniques are respectively. For Arabidopsis Thaliana gene data, UFSC-MOO clustering approach selects five out of features and based on these features the four clusters are evolved (K = 4). The five selected samples are as given in Table 2. The obtained value of Sil(C) is 0.4412 which is found highest when compared to other nine clustering approaches. The cluster count and Sil(C) scores for the clustering techniques are obtained as: MO-fuzzy (4, 0.4162), MOGA (4, 0.4055), FCM (4, 0.3691), SGA (4, 0.3971), Average Linkage (5, 0.3095), SOM (5, 0.2261), CRC (4, 0.4061), Spectral (4, 0.1591) and K-mean (5, 0.3695).

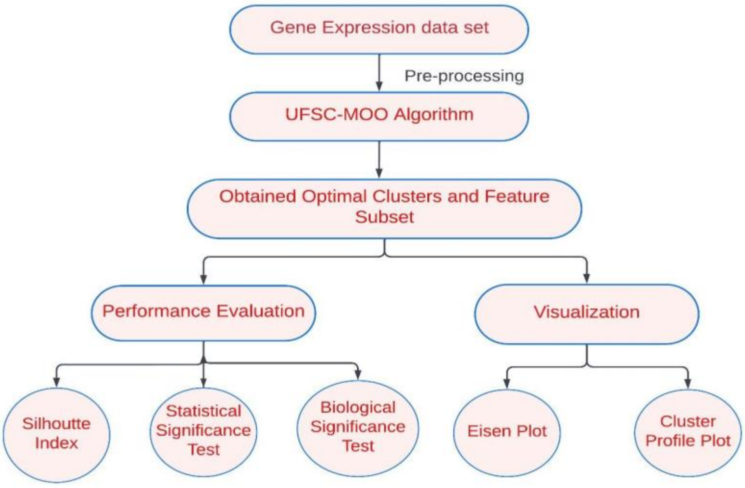


Figure 3. Descriptive View of Result Analysis

Statistical Significance Test

Wilcoxon rank sum test [49] is performed to establish the statistical significance of UFSC-MOO compared to other clustering algorithms. Table 4 displays the p-values, indicating a significance level below 5%. The test compared the Silhouette index medians between UFSC-MOO and other clustering techniques for gene data. The p-value below 0.05 confirms a significant difference supporting the efficacy of UFSC-MOO in gene data clustering.

Table 4. p-values computed for UFSC-MOO in respect to other clustering techniques

Data Set	**MOGA**	**MO-fuzzy**	**SGA**	**SOM**	**FCM**	**K-means**	**Spectral**	**CRC**
Arabidopsis Thaliana	2.55E^{-04}	1.07E^{-03}	3.01E^{-03}	5.78E^{-06}	6.65E^{-05}	3.11E^{-05}	1.25E^{-10}	2.11E^{-03}
Yeast Sporulation	3.62E^{-05}	3.01E^{-05}	3.11E^{-05}	3.79E^{-04}	2.09E^{-08}	3.97E^{-05}	5.97E^{-05}	5.01E^{-05}
Yeast cell cycle	2.54E^{-04}	1.09E^{-04}	2.93E^{-04}	4.44E^{-04}	5.01E^{-06}	2.22E^{-03}	5.01E^{-04}	2.01E^{-03}

The p-values in above table confirm the high statistical significance of Silhouette index values attained by UFSC-MOO.

Biological Significance Test

Gene Ontology annotation [51] database is utilised to embark the biological relevance of the obtained clusters. The probability p is computed to demonstrate the compatibility between number of genes n, for a specific Gene ontology category and cluster of length K. This probability [50] is given by equation below.

$$p = 1 - \sum_{i=0}^{n-1} \frac{\binom{t}{i}\binom{j-t}{K-i}}{\binom{j}{n}} \tag{11}$$

Here, the number of genes for a specific GO group and the total number of genes to genome are represented by t and j respectively. Once the p-value for each GO category is obtained, biological significance test is employed for genes belonging to a cluster. Under any scenario, if p value holds zero, it shows that genes belonging to the particular cluster have identical biological function. In this paper, we conduct the biological significance test for Yeast Sporulation data set at 1% significance level. Moreover, the biological significance test is also conducted for clustering solutions given by different algorithms. The number of clusters for which the most significant GO terms having p-value less than 1% (0.01) in respect of different algorithms are given as- MO-fuzzy (6), MOGA (6), FCM (4), SGA (6), Average Linkage (4), SOM (4), CRC (6), Spectral (6) and K-mean (6). It is observed that for yeast sporulation data set, the count of GO terms for diverse clusters obtained by UFSC-MOO are non-identical, such as- cluster 1 (59 terms), cluster 2 (51 terms), cluster 3 (50 terms), cluster 4 (56 terms), cluster 5 (20 terms) and cluster 6 (28 terms). Now, for MO-fuzzy, the number of GO terms per cluster varies in comparison with UFSC-MOO as- cluster 1 (52 terms), cluster 2 (35 terms), cluster 3 (30 terms), cluster 4 (21 terms), cluster 5 (10 terms) and cluster 6 (49 terms) at 1% significance level. Table 5 represents the p-values of the most significant GO terms of genes of a particular cluster.

Table 5. The three most significant Gene Ontology terms and associated p-values, of six non-identical clusters obtained by UFSC-MOO technique for Yeast Sporulation gene data set.

Clusters Obtained	Significance Gene Ontology term	p-value
Cluster 1	Cytoplasmic translation:GO:0002181	3.36E^{-61}
	Translation: GO:0006412	9.80E^{-32}
	Cellular protein metabolic process: GO:0044267	2.07E^{-17}
Cluster 2	sporulation :GO:0043934	2.95E-39
	anatomical structure formation involved in morphogenesis : GO:0048646	1.47E-38
	sporulation resulting in formation of a cellular spore :GO:0030435	2.2E-38
Cluster 3	reproductive process in single-celled organism:GO:0022413	6.55E-33
	developmental process involved in reproduction:GO:0003006	7.11E-32
	single organism reproductive process :GO:0044702	7.11E-32
Cluster 4	ribosome biogenesis :GO:0042254	1.45E-12
	ribonucleoprotein complex biogenesis :GO:0022613	5.44E-11
	rRNA processing:GO:0006364	4.22E-09
Cluster 5	meiotic nuclear division:GO:0007126	2.90E-26
	meiotic cell cycle:GO:0051321	2.90E-26
	reciprocal DNA recombination :GO:0035825	6.28E-26
Cluster 6	carboxylic acid metabolic process :GO:0019752	5.71E-12
	oxoacid metabolic process :GO:0043436	1.41E-11
	organic acid metabolic process :GO:0006082	5.69E-11

The log transformation of p-values is done to enhance the user readability. The obtained clusters from various clustering algorithms having the significant GO terms are more biologically enriched if $-\log_{10}$ (p value) has higher value (or lower p-value). Figure 4 shows the box-plot representing the six gene clusters. This boxplot below shows the comparison among results obtained by UFSC-MOO, MOGA and MO-fuzzy as all the three methods come up with six numbers of clusters holding the most significant GO-terms associated with their p-values. Now, It can be clearly observed that UFSC-MOO shows higher −log10(p-value) in comparison to MOGA and MO-fuzzy. Therefore, it can be concluded that our proposed UFSC-MOO successfully obtained more biologically and functionally enriched gene clusters.

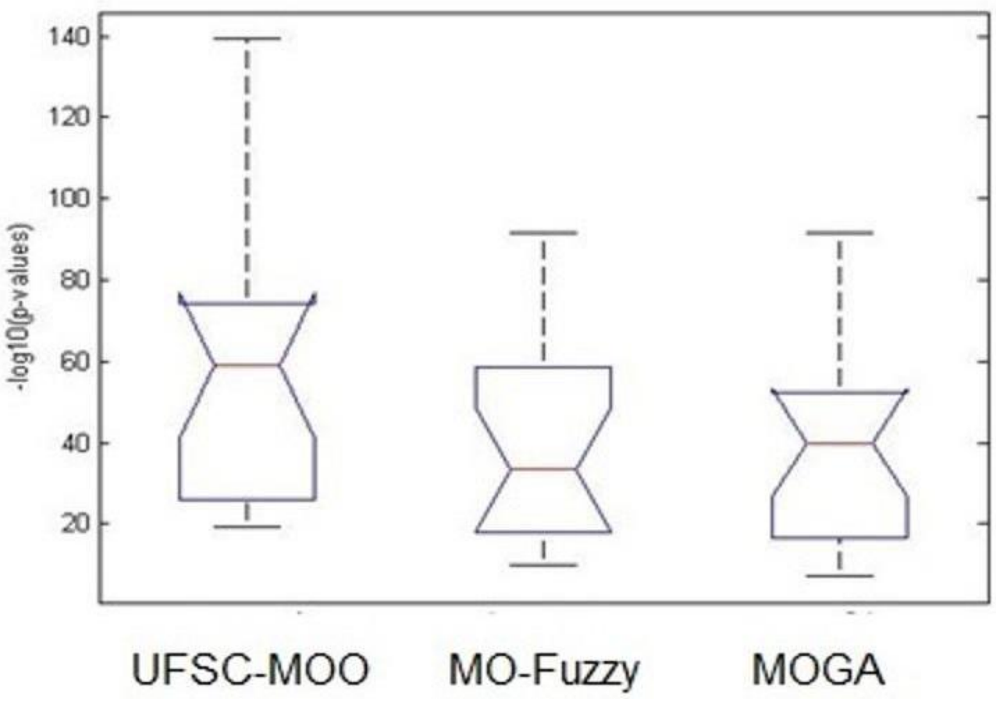


Figure 4. Boxplot of p-values of the most significant GO terms for Yeast Sporulation data set of all the clusters derived by UFSC-MOO, MOGA and MO-fuzzy clustering algorithms

Visual representations of obtained clustering solutions

The obtained clustering solutions are visualised using Eisen plot as shown in Figure 5 and cluster profile plot, given by Figure 6-8 for all three given data sets. Cluster profile plot visualizes gene expression values of resultant clusters over different time points. Prior to plotting, gene expression values are normalized. For cluster profiling, firstly average gene expression values are calculated for each cluster considering different time points. Secondly, within each gene cluster, we compute standard deviation of expression values of different time points. In the end, we plot the cluster's gene expression values with the average and standard deviation represented by a black line. In Eisen plot, the patterns having similar colour shows similar functional behaviour, so they are grouped together. Similarly, genes within same cluster shows similar functional behaviour as their gene expression values hold identical colours. Here, red colour represents higher expression levels. Green colour shows lower expression levels while white colour indicates cluster boundaries. The black colour denotes that differential expression values are not present.

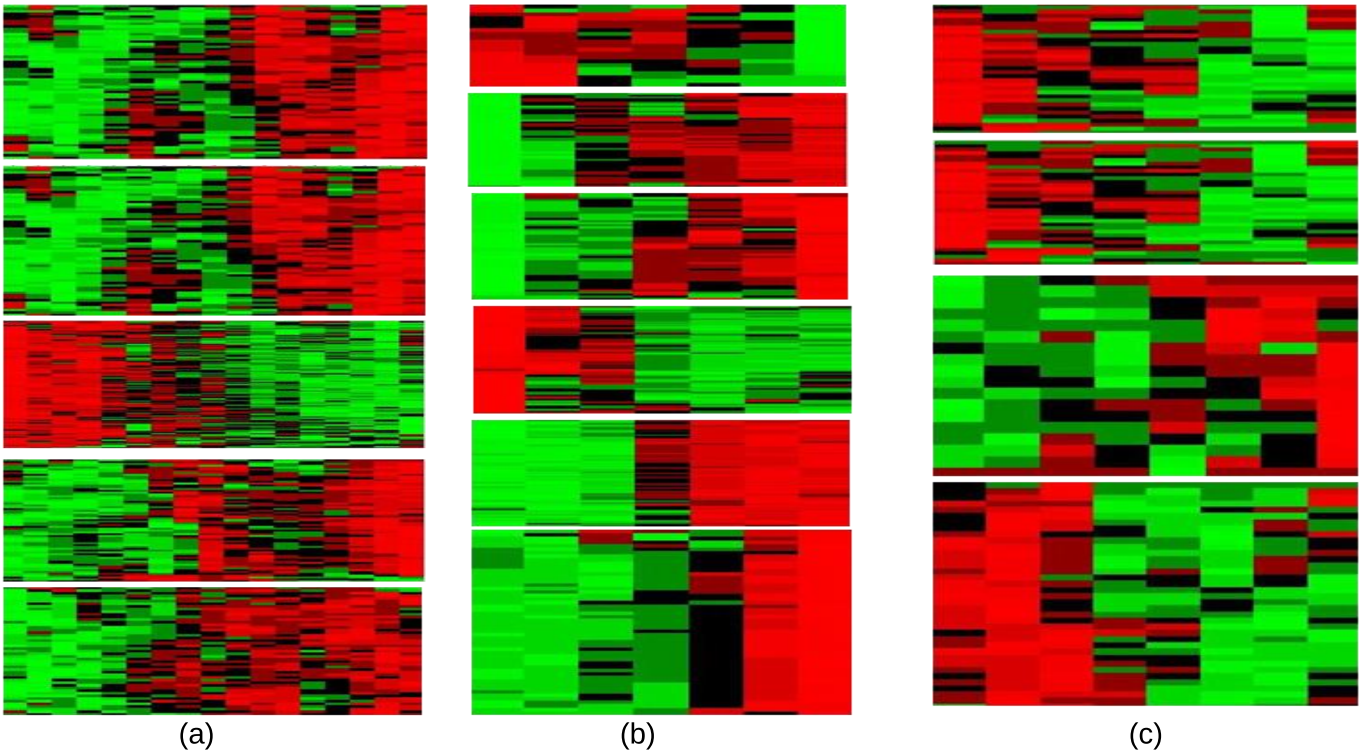
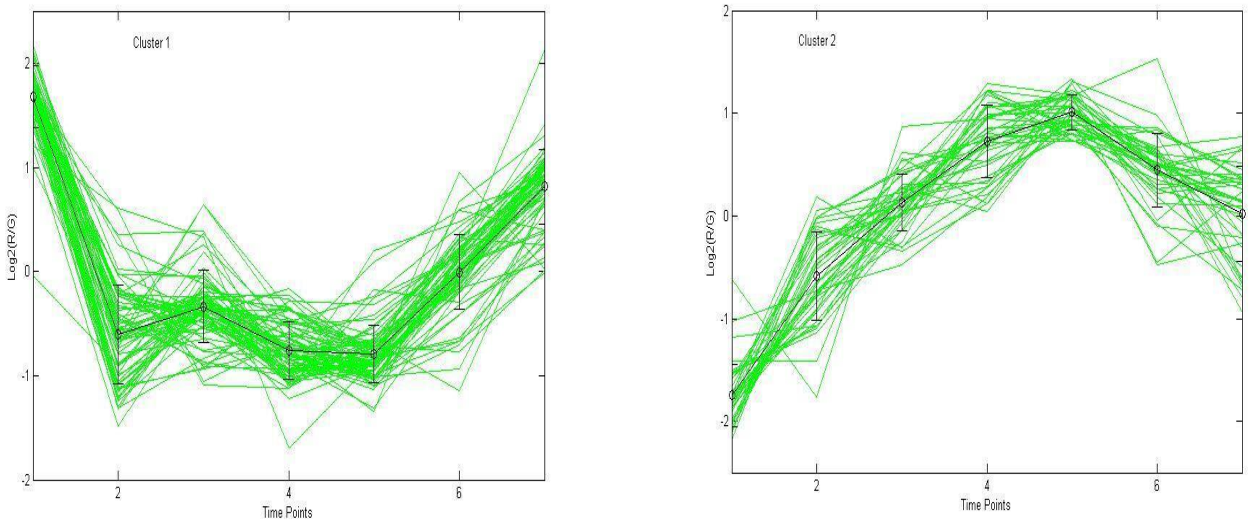


Figure 5. UFSC-MOO generates Eisen plot for gene data clustering in Yeast Sporulation (a), Yeast Cell cycle (b), Arabidopsis Thaliana (c)



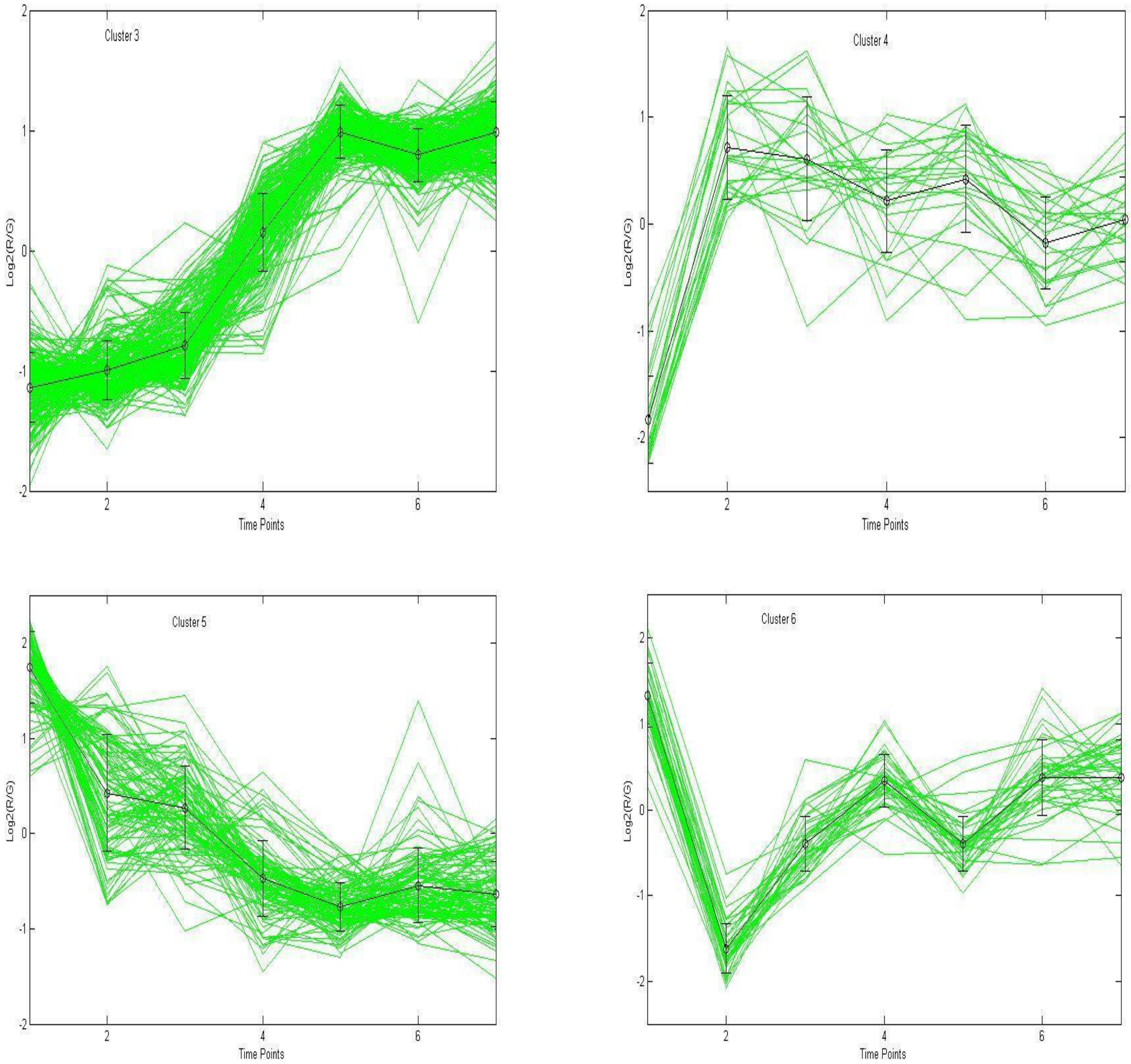


Figure 6. Yeast Sporulation gene data: Cluster Profile Visualization

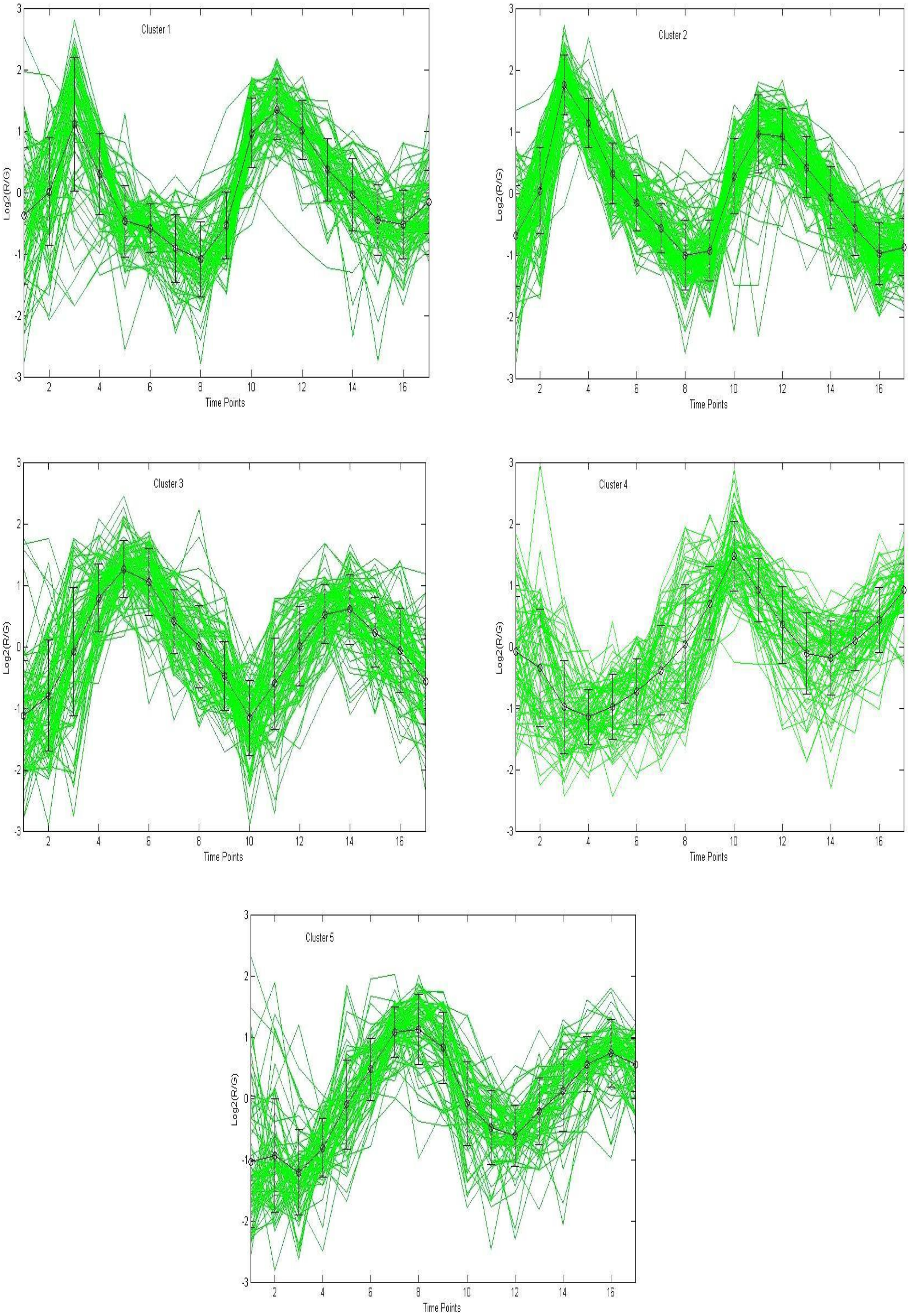


Figure 7. Yeast Cell Cycle data: Cluster Profile Visualization

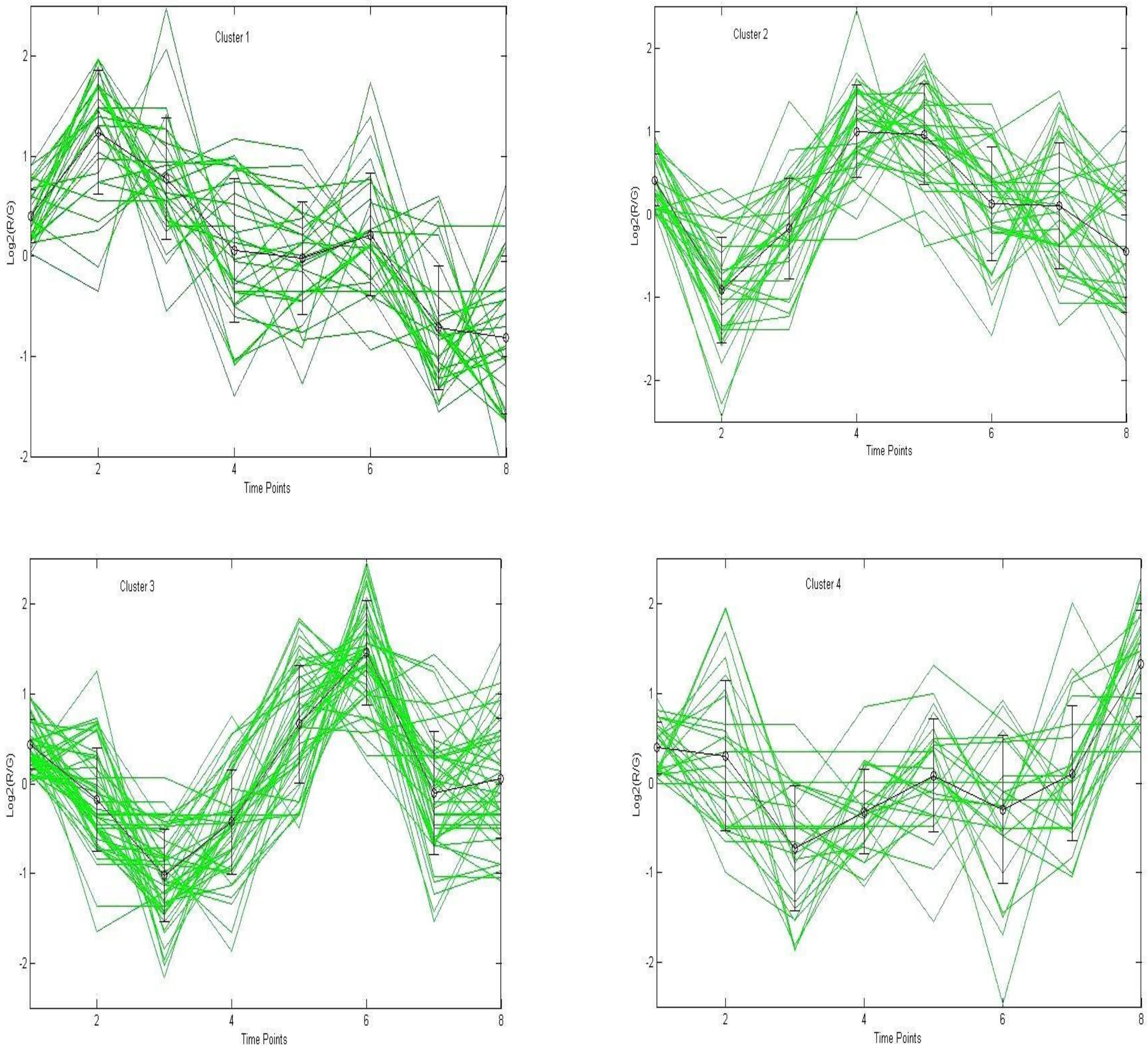


Figure 8. Arabidopsis Thaliana data: Cluster Profile Visualization

The UFSC-MOO technique has demonstrated superior results in simultaneous feature selection and gene data clustering compared to nine other clustering techniques. The inclusion of feature selection in the clustering procedure has significantly reduced the overall computational complexity. Notably, UFSC-MOO outperforms MO-fuzzy [41] and MOGA [42], which are popular multi-objective clustering techniques for gene expression data that lack a feature selection step. Our experimental results highlight the importance of feature selection in gene data clustering. The use of a point symmetry-based distance metric enables the detection of clusters with diverse shapes. Simultaneously optimizing multiple internal validity indices improves cluster partitioning results. Hence, UFSC-MOO effectively identifies suitable cluster centers and feature/sample combinations for gene expression data. The list of used abbreviations is mentioned in Table 6.

Table 6. List of used Abbreviations

Abbreviation	Definition
AMOSA	Archived multi-objective simulated annealing
UFSC-MOO	Unsupervised feature selection and clustering using Multi-objective optimization
Sym index	Symmetry based cluster validity index
MOO	Multi-objective optimization
XB index	Xie-Beni index
FCM	Fuzzy C-means
MO-fuzzy	Multi-objective fuzzy
MOGA	Multi-objective genetic algorithm
SGA	Standard genetic algorithm
SOM	Self organizing map
CRC	Cyclic redundancy check
GO	Gene Ontology
Sil index	Silhoutte index
HL	Hard limit
SL	Soft limit

CONCLUSION AND FUTURE WORK

Our approach, UFSC-MOO, addresses the challenges of multi-objective clustering in gene expression data. We propose a comprehensive framework that simultaneously performs feature selection and unsupervised clustering to identify co-expressed genes. UFSC-MOO optimizes a multi-objective fitness function, combining Sym index, XB index, and feature weight index. By selecting relevant samples and features using feature weight, we reduce the computational complexity of clustering in a reduced dimensional space. Experimental results on three gene expression datasets demonstrate UFSC-MOO's ability to find optimal feature subsets and achieve high-quality partitioning. Comparative analysis with nine clustering methods confirms its superiority. The statistical significance test and biological significance test have proven that obtained clusters are statistically and biologically enriched. In our current research, we acknowledge certain limitations, by considering alternative optimization techniques like NSGA-2, PSO, and differential evolution to evaluate convergence and diversity. Additionally, we are actively exploring refined cluster validity measures to accommodate diverse cluster shapes and sizes. As our approach extends beyond gene expression data, we are dedicated to its application in diverse fields such as cancer data, MR brain image segmentation, and NLP. This expansion encompasses the exploration of advanced initialization techniques, and leveraging supervised information for enhanced clustering outcomes. Future work may include incorporating supervised information and exploring alternative multi-objective approaches for gene data clustering, leveraging semi-supervised learning with ground truth information. Further, we may explore and use several other meta-heuristic techniques for feature selection in gene data.

Funding: “No external funding was received for this research.”
Conflicts of Interest: “No conflict of interest is declared by authors.”

REFERENCES

1. Onan A. Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion. J King Saud Univ Comput Inf Sci. 2023 Jul 7;35(7):101610.
2. Onan A. SRL-ACO:A text augmentation framework based on semantic role labeling and ant colony optimization. J King Saud Univ Comput Inf Sci. 2023 Jul 7;35(7):101611.
3. Onan A, Korukoglu S, Bulut H. Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl. 2016 Mar;57:232-47.
4. Onan A. Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering. IEEE Access. 2019 Oct 7;7:145614-33.
5. Onan A. Biomedical text categorization based on ensemble pruning and optimized topic modelling. Comput Math Methods Med. 2018 Jul 22;2018:2497471.
6. Onan A. An ensemble scheme based on language function analysis and feature engineering for text genre classification. J Inf Sci. 2016 Dec 1;44(1):28–47.
7. Onan A, Korukoglu S, Bulut H. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. Inf Process Manag. 2017 Jul;53(4):814-33.
8. Onan A. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. Concurr Comput. 2021 Dec 10;33(23):e5909.

9. Onan A. Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. Comput Appl Eng Educ. 2021 May;29(3):572-89.
10. Silhavy R, editor. Soft Engineering Methods in Intelligent Algorithms. 8th Computer Science On-line Conference; 2019 Apr 24-27; Prague, Czech Republic. Cham: Springer Nature; 2019 May. 293-304 p.
11. Onan A, Tocoglu MA. A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification. IEEE Access. 2021 Jan;9:7701-22.
12. Onan A. Mining opinions from instructor evaluation reviews: a deep learning approach. Comput Appl Eng Educ. 2020 Jan;28(1):117–38. doi:10.1002/cae.22179.
13. Xue Y, Xue B, Zhang M. Self-Adaptive Particle Swarm Optimization for Large-Scale Feature Selection in Classification. ACM Trans Knowl Discov Data. 2019 Sep 24;13(5):50. doi:10.1145/3340848.
14. Song XF, Zhang Y, Gong DW, Gao XZ. A Fast Hybrid Feature Selection Based on Correlation-Guided Clustering and Particle Swarm Optimization for High-Dimensional Data. IEEE Trans Cybern. 2022 Sep;52(9):9573-86. doi:10.1109/TCYB.2021.3061152. Epub 2022 Aug 18.
15. Zhang Y, Wang YH, Gong DW, Sun XY. Clustering-Guided Particle Swarm Feature Selection Algorithm for High-Dimensional Imbalanced Data with Missing Values. IEEE Trans Evol Comput. 2022 Aug;26(4):616-30. doi:10.1109/TEVC.2021.3106975.
16. Onan A. Consensus clustering-based undersampling approach to imbalanced learning. Sci Program. 2019 Mar 3;2019:5901087. doi:10.1155/2019/5901087.
17. Onan A. Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. J King Saud Univ Comput Inf. 2022 May;34(5):2098-117. doi:10.1016/j.jksuci.2022.02.025.
18. Onan A, Korukoglu S. A feature selection model based on genetic rank aggregation for text sentiment classification. J Inf Sci. 2017 Feb;43(1):25–38. doi:10.1177/0165551515613226.
19. Hancer E. A new multi-objective differential evolution approach for simultaneous clustering and feature selection. Eng Appl Artif Intell. 2020 Jan;87:103307. doi:10.1016/j.engappai.2019.103307.
20. Hancer E, Xue B, Zhang M. A survey on feature selection approaches for clustering. Artif Intell Rev. 2020 Jan 2; 53:4519–4545. doi:10.1007/s10462-019-09800-w.
21. Sahu B, Dehuri S, Jagadev AK. Feature selection model based on clustering and ranking in pipeline for microarray data. Inform Med. 2017 Jul 29;9:107-22. doi:10.1016/j.imu.2017.07.004.
22. Ouadfel S, Elaziz MA. Efficient High-Dimension Feature Selection Based on Enhanced Equilibrium Optimizer. Expert Syst Appl. 2021 Sep 10;187:115882. doi:10.1016/j.eswa.2021.115882.
23. Satapathy SC, Avadhani PS, Abraham A, editors. Advances in Intelligent and Soft Computing. Proceedings of International Conference on Information Systems Design and Intelligent Applications; 2012 Jan 5-7; Visakhapatnam, India. Berlin, Heidelberg: Springer; 2012. 507–514 p.
24. Hancer E. A differential evolution approach for simultaneous clustering and feature selection. Proceedings of the International Conference on Artificial Intelligence and Data Processing; 2018 Sept 28-30; Malatya, Turkey. Piscataway (New Jersey): IEEE; 2019 Jan 24. p. 1–7. doi:10.1109/IDAP43944.2018.
25. Lensen A, Xue B, Zhang M. Using particle swarm optimisation and the silhouette metric to estimate the number of clusters, select features, and perform clustering. In: Squillero G, Sim K, editors. Proceedings of the 20th European Conference on the Applications of Evolutionary Computation; 2017 Apr 19-21; Amsterdam, Netherlands. Champ: Springer; 2017. p. 538–554. doi:10.1007/978-3-319-55849-3_35.
26. Prakash J, Singh PK. Gravitational search algorithm and K-means for simultaneous feature selection and data clustering:a multi-objective approach. Soft Comput. 2019 Mar 1;23(6):2083–100. doi:10.1007/s00500-017-2923-x.
27. Gupta A, Datta S, Das S. Fuzzy clustering to identify clusters at different levels of fuzziness: an evolutionary multiobjective optimization approach. IEEE Trans Cybern. 2021 May;51(5):2601-11. doi:10.1109/TCYB.2019.2907002.
28. Alok AK, Gupta P, Saha S, Sharma V. Simultaneous feature selection and clustering of micro-array and RNA-sequence gene expression data using multiobjective optimization. Int J Mach Learn Cybern. 2020 Jun 1;11(11):2541-2563. doi:10.1007/s13042-020-01139-x.
29. McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE, Engelhardt BE. Clustering gene expression time series data using an infinite gaussian process mixture model. PLoS Comput Biol. 2018 Jan 16;14(1):1-26. doi:10.1101/131151.
30. Mitra S, Saha S. A multiobjective multi-view cluster ensemble technique: application in patient subclassifcation. PLoS ONE. 2019 May 23;14(5):e0216904. doi:10.1371/journal.pone.0216904.
31. Parraga-Alava J, Dorn M, Inostroza-Ponta M. A multiobjective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies. BioData Min. 2018 Aug 7;11(1):16. doi:10.1186/s13040-018-0178-4.
32. Wang Z, Gu H, Zhao M, Li D, Wang J. MSC-CSMC: A multi-objective semi-supervised clustering algorithm based on constraints selection and multi-source constraints for gene expression data. Front Genet. 2023 Feb 27;14:1-13. doi:10.3389/fgene.2023.1135260.
33. Aziz RM. Cuckoo Search-Based Optimization for Cancer Classification: A New Hybrid Approach. J Comput Biol. 2022 Jun;29(6):565-584. doi:10.1089/cmb.2021.0410. Epub 2022 May 6.

34. Aziz RM. Application of nature inspired soft computing techniques for gene selection: a novel frame work for classification of cancer. Soft Comput. 2022 Nov 1;26(12):12179–96. doi:10.1007/s00500-022-07032-9.
35. Aziz RM. Nature-inspired metaheuristics model for gene selection and classification of biomedical microarray data. Med Biol Eng Comput. 2022 Jun;60(6):1627–1646. doi:10.1007/s11517-022-02555-7. Epub 2022 Apr 11.
36. Bandyopadhyay S, Saha S, Maulik U, Deb K. A simulated annealing-based multiobjective optimization algorithm: Amosa. Evolut Comput IEEE Trans. 2008 Jun;12(3):269–83. doi:10.1109/TEVC.2007.900837.
37. Bandyopadhyay S, Saha S. A point symmetry-based clustering technique for automatic evolution of clusters. Knowl Data Eng IEEE Trans. 2008 Nov;20(11):1441–57. doi:10.1109/TKDE.2008.79.
38. Xie XL, Beni G. A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell. 1991 Aug;13(8):841–7.doi:10.1109/34.85677
39. Bandyopadhyay S, Saha S. Gaps: A clustering method using a new point symmetry-based distance measure. Pattern Recogit. 2007 Dec;40(12):3430–51. doi:10.1016/j.patcog.2007.03.026.
40. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. 1st ed. Berlin: Springer;1981 Jan 1.43-95.
41. Saha S, Ekbal A, Gupta K, Bandyopadhyay S. Gene expression data clustering using a multiobjective symmetry based clustering technique. Comput Biol Med. 2013 Nov;43(11):1965–77. doi:10.1016/j.compbiomed.2013.07.021.
42. Bandyopadhyay S, Mukhopadhyay A, Maulik U. An improved algorithm for clustering gene expression data. Bioinformatics. 2007 Nov 1;23(21):2859–65. doi:10.1093/bioinformatics/btm418.
43. Maulik U, Bandyopadhyay S. Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification. Geosci Remote Sens IEEE Trans. 2003 May;41(5):1075–81. doi:10.1109/TGRS.2003.810924.
44. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci. 1999 Mar 16;96(6):2907–12. doi:10.1073/pnas.96.6.2907.
45. Tou JT, Gonzalez RC. Pattern recognition principles. 1st ed. Massachusetts: Addison-Wesley; 1974. 143-229 p.
46. Qin ZS. Clustering microarray gene expression data using weighted chinese restaurant process. Bioinformatics. 2006 Aug 15;22(16):1988–97. doi:10.1093/bioinformatics/btl284.
47. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Lucien M, Cam L, Neyman J, editors. Proceedings of the 5th Berkeley symposium on mathematical statistics and probability; 1965 Dec 27- 1966 Jan 7; Oakland, CA. USA: Euclid; 1967 Jan 1. p. 281–297.
48. Von LU. A tutorial on spectral clustering. Stat Comput. 2007 Nov 01;17(4):395–416.
49. Wilcoxon F, Katti S, Wilcox RA. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. 1st ed. New York: American Cyanamid;1963.25-34 p.
50. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999 Jul; 22(3):281–285. doi:10.1038/10343.
51. Gene Ontolology Term Finder for Yeast Sporulation data set [Internet]. Version 0.86. Stanford (CA): Saccharomyces Genome Database Group. c1997 - [cited 2023 May 15]. Available from: http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.
52. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, et al. The transcriptional program of sporulation in budding yeast. Science. 1998 Oct 23;282(5389):699–705. doi:10.1126/science.282.5389.699.
53. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. Nature. 2001 Jan 25;409(6819):533–8. doi:10.1038/35054095.
54. Li JJ, Huang H, Bickel PJ, Brenner SE. Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by moden code rna-seq data. Genome Res. 2014 Jul;24(7):1086–101. doi:10.1101/gr.170100.113.
55. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987 Nov;20:53–65. doi:10.1016/0377-0427(87)90125-7.
56. Maulik U, Mukhopadhyay A, Bandyopadhyay S. Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes. BMC Bioinform. 2009 Jan 01;10(1):27. doi:10.1186/1471-2105-10-27.