



Simultaneous feature selection and clustering of micro-array and RNA-sequence gene expression data using multiobjective optimization

Abhay Kumar Alok¹ · Pooja Gupta² · Sriparna Saha¹ · Vineet Sharma²

Received: 22 June 2019 / Accepted: 2 May 2020 / Published online: 1 June 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

In this paper, we have devised a multiobjective optimization solution framework for solving the problem of gene expression data clustering in reduced feature space. Here clustering problem is viewed from two different aspects: clustering of genes in reduced sample space or clustering of samples in reduced gene space. Three objective functions: two internal cluster validity indices and the count on the number of features are optimized simultaneously by a popular multiobjective simulated annealing based approach, namely AMOSA. Here, point symmetry based distance is used for the assignment of gene data points to different clusters. Seven publicly available benchmark gene expression data sets are used for experimental purpose. Both aspects of clustering in reduced feature space is demonstrated. The proposed gene expression clustering technique outperforms the existing nine clustering techniques. Apart from this, also some statistical and biological significant tests have been carried out to show that the proposed FSC-MOO technique is more statistically and biologically enriched

Keywords Gene expression data clustering · Feature selection · Point symmetry based distance · Multiobjective optimization · Cluster validity index

1 Introduction

A large matrix showing gene expression levels(rows) and the different experimental conditions(columns) represent Gene expression data. Clustering of gene expression data

can be carried out in two different spaces: gene space or sample space [24, 25, 31, 32]. In [11, 23], it has been mentioned that the appropriate sample selection helps to get a low-level visual representation of gene behavior across the samples. This dimensionality reduction in sample space helps to effectively tackle the problem of determining a low dimensional embedding that provides a precise visual representation of gene-gene interactions. Inspired by this observation in [23], a feature selection technique is proposed to reduce the number of samples from a given gene expression data set. The identified co-expressed genes are highly symmetrical, overlapping, and high-dimensional in nature. Most of the single-objective based clustering techniques fail to evolve the optimal clusters. Recently, some authors have developed multiobjective based clustering techniques [18, 19, 21, 22] and applied on gene expression data to obtain the true partitioning results. But none of these authors have tried to solve simultaneous feature selection and clustering for gene expression. Thus none of these techniques have focussed on developing a system which is responsible for simultaneous feature selection and clustering.

In this paper, we have posed the problem of sample/gene selection for the clustering of genes/samples as a

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13042-020-01139-x>) contains supplementary material, which is available to authorized users.

✉ Abhay Kumar Alok
abhayalok@iitp.ac.in

Pooja Gupta
pooja.gupta@kiet.edu

Sriparna Saha
sriparna@iitp.ac.in

Vineet Sharma
vineet.sharma@kiet.edu

¹ Computer Science Engineering, Indian Institute of Technology, Patna, India

² Computer Science Engineering, Krishna Institute of Engineering and Technology, AKTU, Ghaziabad, Lucknow, India

multiobjective optimization problem (MOOP). Here, AMOSA [4] a simulated annealing based optimization technique is used for the optimization of multiple objectives. The proposed FSC-MOO, a feature selection as well as multiobjective clustering technique, encodes the number of features (in this case, either samples/genes) and the number of cluster centers (in terms of genes/ samples) in the form of a solution. Here, to obtain the true partitioning of genes/samples, we have employed a newly developed point-symmetry based distance [1]. The three objective functions are simultaneously optimized to obtain optimum clustering solutions. The First two internal cluster validity indices used here, explore different data intrinsic properties. The measured validity indices are Sym-index [2] and XB-index [37]. Sym-index [2] is responsible for counting total symmetrical compactness within the clusters, whereas XB-index [37] uses euclidean distance to measure cluster separation and cluster compactness. The third objective function is the feature index that is responsible for extracting the relevant features.

Now with the help of AMOSA, these two cluster validity measures and the number of samples/genes are simultaneously optimized to evolve the final non-dominated solutions. These non-dominated solutions represent different cluster centers as well as multiple combinations of samples/genes. The algorithm will automatically identify the appropriate set of samples in case of gene clustering/appropriate set of genes in case of sample clustering, several gene-clusters/sample-clusters, and the proper partitioning of genes/samples from a given gene expression data set.

The performance of the proposed technique (FSC-MOO) is demonstrated on publicly available seven real-life gene expression data sets. Also, to show the effectiveness of the proposed clustering algorithm FSC-MOO, it is compared with well known existing clustering techniques. These techniques include MOGA clustering [3], MO-fuzzy [29], FCM [5], SGA [16], Hierarchical average linkage clustering [34], Self Organizing Map(SOM) [30], Chinese Restaurant Clustering(CRC) [26], Spectral Clustering [35] and K-mean clustering [14]. Further, some statistical and biological significant tests are performed. The biological significance test reveals that indeed the appropriate sample selection helps in forming the biologically enriched gene clusters. The performance of the proposed sample clustering technique, which simultaneously performs sample clustering and gene selection, is shown for two other gene expression data sets. For the comparative analysis, we have used MO-fuzzy, FCM, Average Linkage, SGA, Spectral, and K-mean clustering techniques. The obtained results show the superiority of the FSC-MOO technique. List of abbreviation is shown in Table 1 to understand the used symbol.

Table 1 List of abbreviations and symbols used

S(C)	Silhouette score/index
ARI	Adjusted rand index
Sym(K)	Validity measure based on point symmetry property
XB	Xie and Beni proposed cluster validity measure
K	Number of clusters
MOO	Multiobjective optimization
MOOP	Multiobjective optimization problem
CA	Classification Accuracy
DB	Davies- Bouldin index
GO	Gene ontology
d_{ps}	Point symmetry based distance
f	Number of features present in a string
MOGA	Multiobjective genetic algorithm
MO-fuzzy	Multiobjective based fuzzy clustering technique
SA	Simulated annealing

2 Problem definition

Let us consider the data matrix $Y = \{\bar{y}_j : j = 1, 2, \dots, n\}$, \bar{y}_j is a vector of dimension D , which is required to be partitioned into K different clusters such that $\sum_{k=1}^K \sum_{j=1}^n u_{kj} = n$, where u_{kj} denotes the membership value of j th point for k th cluster. Not all features are important from clustering point of view. So, here the goal is to project the original D -dimensional data into F -dimensional subspace such that $F \leq D$. Also, to simultaneously satisfy multiple cluster quality measures (as opposed to single-objective optimization), the problem of clustering is posed as a multiobjective optimization problem (MOOPS). Multiobjective optimization problems can be formulated as: select the most efficient vectors from the given decision variables $\bar{y}^* = [y_1^*, y_2^*, \dots, y_n^*]^T$ which are bound to satisfy n inequality constraints:

$$h_i(\bar{y}) \geq 0, \quad i = 1, 2, \dots, n, \quad (1)$$

the q equality constraints

$$g_i(\bar{y}) = 0, \quad i = 1, 2, \dots, q, \quad (2)$$

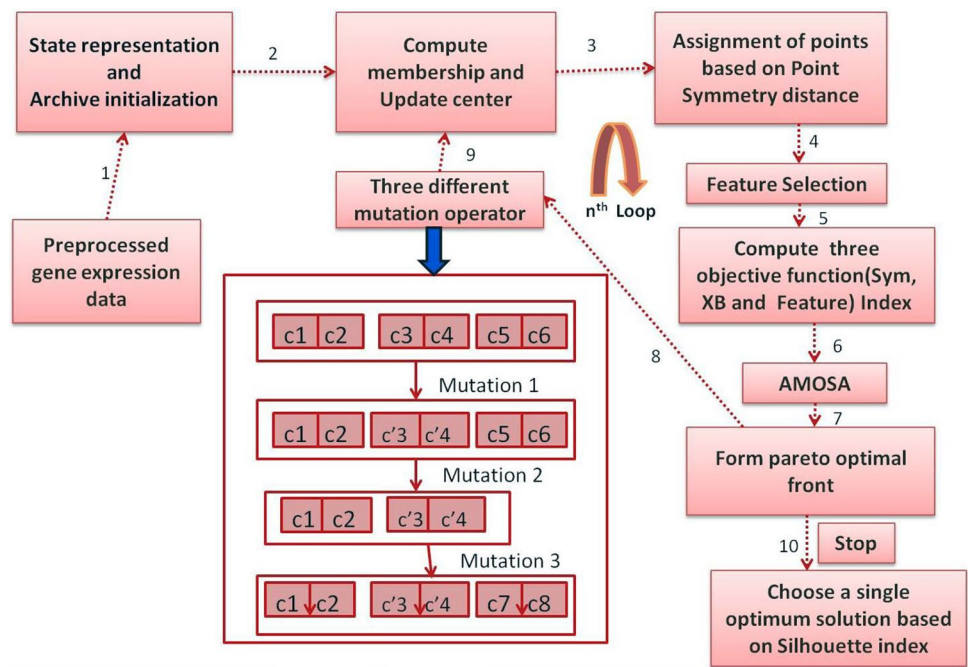
and also simultaneously optimize P objective functions

$$F_1(\bar{y}), F_2(\bar{y}), \dots, F_P(\bar{y}). \quad (3)$$

The mathematical formulation of the proposed problem is given below:

1. Input: A set of points $Y = \{\bar{y}_j : j = 1, 2, \dots, n\}$, where \bar{y}_j is a vector of dimension D . Here n denotes the number of data points.

Fig. 1 Working principle of our proposed multiobjective clustering algorithm



2. Output: Determine a subset of feature combinations F where $F \leq D$ and based on this subset of features partition the data into some automatic number of clusters (K). The value of K would be determined automatically. The partition matrix U of size $K \times n$ is generated where K : number of clusters and n : number of points such that $\sum_{k=1}^K \sum_{j=1}^n u_{kj} = n$, where u_{kj} denotes the membership value of j th point for k th cluster. $u_{kj} = 0$ if j th point does not belong to k th cluster else $u_{kj} = 1$ if j th point belongs to k th cluster.
3. Optimization functions: Several cluster validity indices like $Val_1, Val_2, \dots, Val_p$ where Val_i denotes i th cluster validity index should be simultaneously optimized using the search capability of any MOO based technique.

3 Proposed method for multiobjective feature selection and unsupervised clustering technique

In this section, we have elaborated the working principle of the proposed FSC-MOO technique in detail. The steps involved in the proposed algorithm FSC-MOO are shown in Fig 1. Here AMOSA [4], an optimization technique based on the properties of simulated annealing, is used as the underlying multiobjective optimization technique to optimize the objective functions simultaneously. In case of multiobjective optimization, a set of trade-off solutions is obtained in which some are dominating and non-dominating to one another. The non-dominating solutions are Pareto optimal set. An example of dominance and non-dominance is shown

in Fig. 2. It can be seen from this figure that there are five solutions in total, and out of these five solutions, four solutions are non-dominating to each other, and these three solutions dominate the one solution.

3.1 Encoding of strings and initialization of archive

In FSC-MOO, the state representation of AMOSA consisted of two different elements. The first one exhibits a set of real numbers that are in fact, the coordinates of centers of partitions of data set. The second one exhibits a set of binary numbers that explore the different sample/gene combinations. With the help of this encoding scheme, AMOSA is capable of evolving true partitioning represented by cluster centers as well as suitable sample/gene combinations.

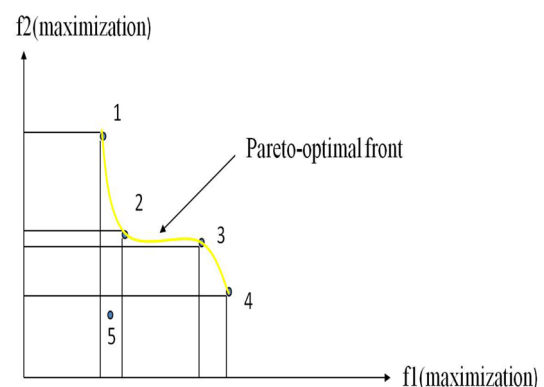
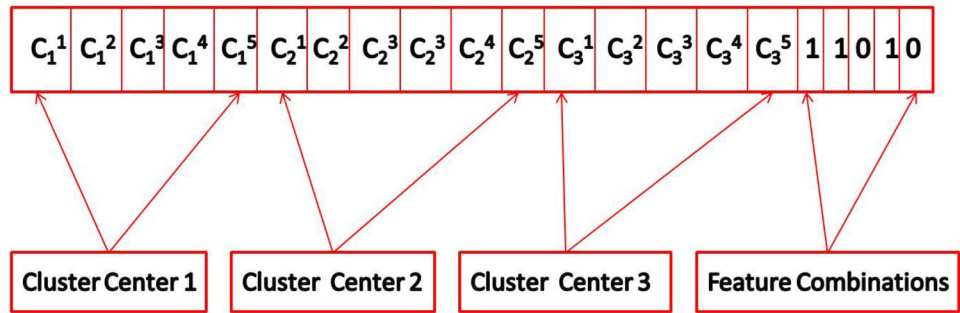


Fig. 2 Pareto-optimal Front and dominance relation

Fig. 3 Representation of cluster centers and features in the form of a string



Let in a particular string total number of features and encoded cluster centers be F and K , respectively. Therefore, the length of the string can be defined as $F + K \times F$. Here randomly chosen binary strings are represented as feature combinations.

Let us assume that in a string, we have $K=3$ number of clusters and $F=5$ number of features. Then the corresponding representation is mentioned below:

$\langle c_1^1, c_1^2, c_1^3, \dots, c_1^5, c_2^1, \dots, c_2^5, c_3^1, c_3^2, \dots, c_3^5, 11010 \rangle$. Here $\langle c_1^1, c_1^2, c_1^3, \dots, c_1^5 \rangle$, $\langle c_2^1, \dots, c_2^5 \rangle$, and $\langle c_3^1, c_3^2, \dots, c_3^5 \rangle$ represent three cluster centers for a single partitioning and only first, second and fourth features are selected. These features are considered for the computation of distance functions for cluster assignment and objective function calculations. The encoding of cluster centers and feature combinations is shown in Fig. 3.

Initially, in the archive each string j encompasses K_j number of clusters. Here value of K_j is generated as per the following formula: $K_j = (rand() \times (K^{max} - 1)) + 2$. Here $rand()$ function is used to generate a random number, which returns an integer number. K^{max} is the maximum allowable cluster number for the data set. Thus in the above Equation after performing $\%$ operation, we can get values in the range of 0 to $(K^{max} - 2)$. Thus by adding 2 to each of the numbers in the given range will return values in the range 2 to K^{max} .

Here we have adopted a random procedure for initialization purposes. Let K_j represent the number of cluster centers for a particular chosen string j . These K_j cluster centers are randomly chosen from the whole data set. Thereafter, these selected cluster centers are encoded in a particular chosen j th string. Also, the minimum center-based distance criterion is used to generate the initial partitionings. Apart from this, a random procedure is also invoked for the initialization of features associated with a particular string. Now let us assume there are total F number of features in a data set; then randomly each position of the feature set is initialized to either 0 or 1. If at i th position, 0 value has occurred, then the feature at that particular i th position is totally inactive for participation in further processing. Similarly, 1 value

indicates that the feature at that particular i th position is participating actively in further processing. Here, further processing specifies the cluster assignment and computation of objective function values etc.

3.2 Assignment of points

Here for the assignment of data points to different clusters, point symmetry based distance $d_{ps}(\bar{y}, \bar{z})$ [1] is utilized. In the proposed technique FSC-MOO, we have considered cluster centers as representatives of different clusters. With this assumption, the assignment of points is done with respect to different clusters. Now let us suppose for a particular string: total number of present features= F and number of clusters= K . Now, the assignment is done using the following steps:

- A gene \bar{x}_j , $1 \leq j \leq n$, is assigned to cluster k using the point symmetry based distance if the following two conditions are satisfied:
 - $d_{ps}(\bar{x}_j, \bar{c}_k) \leq d_{ps}(\bar{x}_j, \bar{c}_i)$, $i = 1, \dots, K, i \neq k$,
 - $d_{sym}(\bar{x}_j, \bar{c}_k) = (d_{ps}(\bar{x}_j, \bar{c}_k) / d_e(\bar{x}_j, \bar{c}_k)) \leq \theta$.
- In case of $(d_{ps}(\bar{x}_j, \bar{c}_k) / d_e(\bar{x}_j, \bar{c}_k)) > \theta$, point \bar{x}_j is assigned to some cluster m using the Euclidean distance as done in case of K-means algorithm. The value of m is determined, which satisfies the following condition: $d_e(\bar{x}_j, \bar{c}_m) \leq d_e(\bar{x}_j, \bar{c}_i)$, $i = 1, 2 \dots K, i \neq m$.

Here, d_{sym} computes the symmetry of a point within a cluster with respect to the cluster center. If the value of this measurement is small enough, then we can expect that the given point is indeed symmetrical with respect to the cluster center. Thus we have used a threshold value to check the amount of symmetricity. If the amount of symmetricity is greater than the threshold, then it signifies that the point lacks symmetry with respect to any cluster. In such cases, the use of Euclidean distance for cluster assignment as done in

case of K-means algorithm is justified. The reasons for this action are as follows: sometimes during the initial stage of the algorithm, clusters may not be appropriately formed. In such situations, the minimum value of d_{ps} for different points is expected to be high due to the lack of point symmetry property with respect to various cluster centers. So, in such conditions, Euclidean distance is used for cluster assignment. In other words, if d_{ps} value is small, then the cluster assignment is done based on symmetry property.

Here in the proposed technique FSC-MOO, the distance computations are done only using those features which are present in that particular string. In this regard, the value of θ is kept as the maximum nearest neighbor distance among all the points for a particular data set.

3.3 Objective functions used

The objectives are selected, keeping the following points in mind:

1. Three different objective functions explore different data intrinsic properties.
2. Sym-index basically measures the total symmetrical compactness of the data. It tries to capture the clusters which are highly symmetric about their cluster centers.
3. XB-index measures the compactness and separation of the obtained clusters using popular Euclidean distance. It tries to detect clusters which are hyperspherical in shape.
4. The last feature index is responsible for extracting the relevant features. The above two objective functions use Euclidean distance in their operations. Sym-index is based on point symmetry-based distance, which again uses Euclidean distance for its computation. Thus these indices are biased towards low-dimensions (low feature values). These indices try to decrease the number of features present in the data set. The optimal values of these indices are obtained at number of features=1. In order to get rid of this bias, this objective function is used which tries to increase the number of features.

To evolve the good clustering solution and the appropriate feature combination, three different objective functions are used, which are simultaneously optimized by the search capability of AMOSA [4].

3.3.1 Sym-index: symmetry based cluster validation index

Sym-index [2] is based on the point-symmetry based distance [1]. It computes the total symmetry present in the partitioning with respect to cluster centers. Let us consider the data matrix $Y = \{\bar{y}_j : j = 1, 2, \dots, n\}$ which is partitioned into

K different clusters. Computation of cluster center \bar{z}_i is done as follows: $\bar{z}_i = \frac{\sum_{j=1}^{n_i} \bar{y}_j^i}{n_i}$.

$$Sym(K) = \left(\frac{1}{K} \times \frac{1}{\mathcal{E}_K} \times D_K \right) \tag{4}$$

Here , $\mathcal{E}_K = \sum_{i=1}^K E_i$, $E_i = \sum_{j=1}^{n_i} d_{ps}^*(\bar{y}_j^i, \bar{z}_i)$,
 $D_K = \max_{i,j=1}^K \|\bar{z}_i - \bar{z}_j\|$.

$$d_{ps}(\bar{y}, \bar{z}) = d_{sym}(\bar{y}, \bar{z}) \times d_e(\bar{y}, \bar{z}), \tag{5}$$

$$= \frac{\sum_{i=1}^{knear} d_i}{knear} \times d_e(\bar{y}, \bar{z}), \tag{6}$$

Among all the cluster center pairs, the maximum separation between any two centers is denoted by D_K . Separation is measured in terms of the Euclidean distance. With some constraint $d_{ps}^*(\bar{y}_j^i, \bar{z}_i)$ is computed according to Equation (5). Here, $knear$ number of nearest neighbors of $\bar{y}_j^i = 2 \times \bar{z}_i - \bar{y}_j^i$ are determined from the points which are part of the cluster i . To identify the actual number of clusters, and to avoid overlapping, Sym-index should be maximized. The different components of Sym-index and their interactions are explained well in [2].

3.3.2 XB index

Xie and Beni [37] focused on the compactness and separability of clusters to define the XB-index. From Equation 7, it can be easily seen that XB-index is the ratio of cluster compactness and cluster separation. For optimal partitioning, the compactness of the clusters should be minimized, and separation between different clusters should be maximized. Hence, the most desirable partitioning is obtained by minimizing the XB-index for $K = 1, \dots, K_{max}$.

$$XB = \frac{\sum_{i=1}^K \sum_{j=1}^n \mu_{ij}^2 \|\bar{x}_j - \bar{c}_i\|^2}{n(\min_{i \neq k} \|\bar{c}_i - \bar{c}_k\|^2)} \tag{7}$$

where, K = Total number of clusters present in a solution. n = Total number of data points to be clustered. $\mu_{ij} = 1$ if j th data point belongs to i th cluster. $\mu_{ij} = 0$ if j th data point does not belong to i th cluster. $x_j = j$ th data point. $c_i = i$ th cluster.

3.3.3 Feature selection based objective function

The third objective function is the number of features (samples or genes) encoded in a particular string. We have to calculate the number of features (samples or genes) present in a particular string, and have to maximize the number of features. This objective function is used to balance the bias of the first two objective functions on dimensionality. Internal cluster validation techniques are based on some distance

computations and thus, those are biased towards lower dimensions [9]. To balance this bias, we have used the third objective, which will try to increase the number of features (samples or genes) present in a data set. $f_3 = \text{maximize } \|f\|$ where $\|f\| =$ number of features present in that particular string. The values of the above mentioned two cluster validity indices are calculated for the partitioning represented in each string. For Sym-index and XB-index computations, we have to consider only those features (samples or genes) which are present in that particular string. The multiple objectives corresponding to a particular string are:

$$obj = \{Sym(K, f), 1/XB(K, f), \|f\|\}$$

where $Sym(K, f)$, $XB(K, f)$, and $\|f\|$ denote the obtained values for Sym-index, XB-index, and the number of features respectively, present in that particular string. The number of clusters encoded in a particular string is denoted by K , and f is the number of features (samples or genes) present in that particular string. The modern MOO technique AMOSA is deployed to maximize these three objectives simultaneously.

3.4 Mutation operations

The intuition behind the mutation operation is to generate a new string from the current string. For this, we have employed the three different types of mutation operations to change the cluster centers encoded in a particular string. Here, the binary mutation is used in which each bit position of feature-combination in a particular string, is flipped with some probability. In other words, if the initial bit position of feature is 1, it can be replaced by 0; and vice versa. The proposed approach is used to select the number of gene-clusters/sample-clusters from any gene expression data set automatically. We initialize each string with some randomly generated number of clusters. To explore the entire search space, we need to change the number of clusters encoded in a particular string. Mutation operations are defined in such a way so that these can change the number of clusters encoded in a particular string. For this, we have developed the second and third mutation operations. The third mutation provides a way to increment the number of clusters. The second mutation operation provides a way to decrement the value of the number of clusters encoded in a particular string. The first mutation operator is used to change the existing cluster centers by some value.

1. Here Laplacian distribution $p(\epsilon) \propto e^{-\frac{|\epsilon-\mu|}{\delta}}$ is used to perturb the individual cluster centers, where δ is the scaling factor which determines the perturbation value. The position for perturbation is denoted by μ , and scaling factor δ is chosen to one. In this way, the old value is replaced by the newly generated value. Independently this perturbation operation is imposed on all the dimen-

sions. Here, the binary mutation is utilized to change the feature combination.

2. To reduce the cluster numbers for a particular string, this mutation operation is utilized. It is done after deleting a randomly selected cluster center from a particular string. Again to change the feature combination, binary mutation is employed.
3. To increase the cluster numbers for a particular string, this mutation operation is utilized. It is done after addition of a randomly selected point from the entire data set in the particular string as a new cluster center. Again to change the feature combination, binary mutation is employed.

We perform any one of these mutation operations on a string if it is selected for mutation process. All these three operations are applied with a uniform probability.

3.5 Selection of the best solution

After the application of the proposed MOO based technique, a set of nondominated solutions on the final Pareto optimal front is obtained [8]. Each of these solutions provides a combination of feature-collection and cluster centers. Based on this information using the point symmetry based distance [1], we can get the partitioning of the gene expression data associated with this solution. All the nondominated solutions are important. But sometimes depending on user requirements and also for the purpose of comparison, we may need to select a single solution. In this connection, we use an internal cluster validity index, Silhouette index [28]. The range of $S(C)$ index value is $[-1, +1]$. Therefore, high positive value of $S(C)$ index ensures the true optimal clustering solution.

4 Data sets and results analysis

Here, seven benchmark gene expression data sets are used for experimental purposes. Two variations of these data sets are used, one variation with preprocessing steps and the other without using any preprocessing steps. The motivation behind these variations is to show the impact of preprocessing steps on the model performance. Used data sets are Yeast Sporulation [6]¹, Yeast Cell Cycle [12]², Arabidopsis Thaliana [27]³,

¹ <http://cmgm.stanford.edu/pbrown/sporulation,sps25>

² <http://faculty.washington.edu/kayee/cluster,sps26>

³ <http://homes.esat.kuleuven.be/thijs/Work/Clustering.html.sps27>

Table 2 Description of data sets where n and d denote the number of genes and the number of samples respectively, used for gene-clustering

Data set	n	d
Yeast Sporulation	474	7
Yeast Cell Cycle	384	17
Arabidopsis Thaliana	138	8
D. melanogaster	14300	30
C. elegans	27135	35
SRBCT	200	72
Leukemia	200	72

D. melano-gaster [13]⁴, C.elegans⁵, SRBCT⁶ and Leukemia⁷. Preprocessed data information are shown in Table 2. The entire descriptive data analysis is mentioned in the given supplementary file.

4.1 Performance metrics

Here, Silhouette index [28] and DB index [7] are used to measure the performance of the proposed FSC-MOO clustering technique. Also, cluster profile [17] and Eisen plots [17] are used to show the obtained partitioning results visually. Also, some biological significance test has been performed to confirm that the obtained partitioning results are more biologically enriched. Apart from this, for SRBCT and Leukemia data sets, two performance measures, namely Adjusted Rand Index [10] and Percentage Classification Accuracy(%CA) [20] are utilized to compare the obtained partitioning results with other clustering algorithms. Again, for the proposed clustering technique FSC-MOO, a statistical significance test is performed using Wilcoxon's rank sum test [36]. The detail information about these performance indicators is mentioned in the given supplementary file.

4.2 Results analysis

In FSC-MOO, a modern MOO technique based on the concepts of simulated annealing, namely AMOSA, is used for simultaneous feature selection and unsupervised clustering. Experimental results are summarized in two different categories: first the proposed technique is applied on five gene-expression data sets for simultaneous sample selection and gene-clustering. In the second part, the proposed technique is applied on two gene-expression data sets for simultaneous

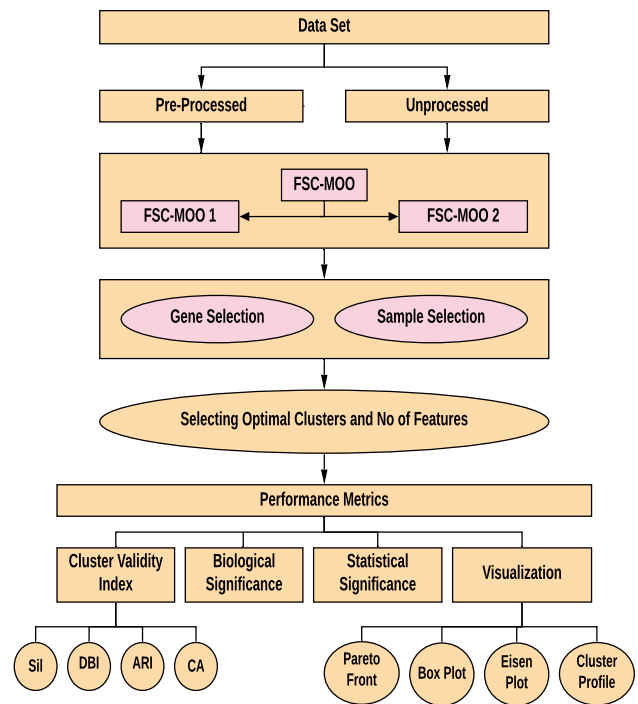


Fig. 4 Descriptive view of result analysis

gene selection and sample clustering. In both the cases we have compared the performance of the proposed approach with some bench-mark approaches. The whole experimental steps are shown in Fig. 4.

The parameters of the proposed FSC-MOO clustering technique are as follows: $SL=100$ $HL=50$, $iter=50$, $Tmax=100$, $Tmin=0.00001$ and cooling rate, $\alpha = 0.9$. The performance of FSC-MOO technique is compared with MO-fuzzy [29], MOGA [3], FCM [5], SGA [16], Average Linkage [34], SOM [30], and CRC [26], Spectral [35] and K-mean [14] clustering techniques where all the time points are utilized for distance computations for the purpose of assignment of genes and objective function calculations. The number of clusters and the number of time points automatically determined by the proposed FSC-MOO technique using point symmetry-based distance are reported in Table 3. To evaluate the quality of the obtained partitioning by the proposed FSC-MOO clustering technique, Silhouette index and DB index values are calculated. As maximum value of Silhouette index and low value of DB index value refer to good partitioning solutions. The Silhouette index and DB index values for optimal solutions generated by different clustering techniques are shown in Table 4. We have also performed two variants of FSC-MOO technique based on varying objective functions.

The parameters of the these two versions of FSC-MOO clustering technique are as follows: $SL=100$ $HL=50$,

⁴ <http://www.stat.ucla.edu/jingyi.li/software-and-data.html,mode>

⁵ <http://www.stat.ucla.edu/jingyi.li/software-and-data.html,mode>

⁶ <http://www.biobab.si/supp/bi-cancer/projections/info/SRBCT.htm>

⁷ <http://www.biobab.si/supp/bi-cancer/projections/info/leukemia.htm>

Table 3 Time points selected, number of clusters and Silhouette index of the final partitionings automatically determined by the proposed FSC-MOO clustering algorithm. Here K denotes number of clusters automatically selected by FSC-MOO clustering algorithm, raw denotes unprocessed data and pre denotes preprocessed data

Data set	Feature selected	K	S(C)
Yeast Sporulation	1,2,3,6,7	6	0.6212
Yeast Cell Cycle	1,2,3,4,5,6,8,9,17	5	0.4421
Arabidopsis Thaliana	1,2,4,6,8	4	0.4258
Yeast Sporulation(raw)	1,3,5,6,7	6	0.5317
Yeast Cell Cycle(raw)	2,5,7,8,10,11,14,15	5	0.2409
Arabidopsis Thaliana(raw)	1,3,4,5,6,8	4	0.5983
D. melanogaster(raw)	1,5,7,8,11,12,14,17,19,22,27,29	3	0.8908
C. elegans(raw)	5,7,8,11,14,15,17,18,20,22,27,29,31	3	0.8897
D. melanogaster(pre)	1,2,4,6,8,9,12,16,22,23,27,30	3	0.9081
C. elegans(pre)	1,4,6,8,10,14,15,17,19,22,25,28,30	7	0.9654

Table 4 Obtained Silhouette index and DB index values for different clustering techniques

Algorithm_used	Sporulation			Cell Cycle			Thaliana		
	K	S(c)	DB	K	S(c)	DB	K	S(c)	DB
FSC-MOO	6	0.6212	0.2362	5	0.4421	0.2467	4	0.4258	0.1469
FSC-MOO ₁	6	0.5944	0.2441	5	0.4413	0.2506	4	0.3958	0.2204
FSC-MOO ₂	6	0.6084	0.2415	5	0.4171	0.2924	4	0.4204	0.1692
MO-fuzzy	6	0.5877	0.2592	5	0.4342	0.2600	4	0.4194	0.1802
MOGA	6	0.5754	0.2851	5	0.4232	0.2889	4	0.4023	0.2678
FCM	7	0.4696	0.4178	6	0.3856	0.4017	4	0.3665	0.4234
SGA	6	0.5712	0.3103	5	0.4232	0.2904	4	0.3854	0.3982
Average Linkage	6	0.5023	0.3925	4	0.4378	0.5206	5	0.3162	0.5124
SOM	6	0.5794	0.2982	6	0.3862	0.4248	5	0.2352	0.6246
CRC	8	0.5623	0.3262	5	0.4275	0.2762	4	0.3965	0.4146
Spectral Clustering	6	0.5606	0.3428	4	0.3626	0.4423	4	0.1682	0.5755
K-mean	6	0.4677	0.4456	5	0.4167	0.4232	5	0.3706	0.4168

Bold symbol indicates high index value obtained by proposed algorithm

$iter=50$, $Tmax=100$, $Tmin=0.00001$ and cooling rate, $\alpha = 0.9$ These two versions are:

- *FSC-MOO₁*: FSC-MOO clustering technique optimizing FCM-index [5] as internal cluster validity indices and the number of features as another objective function.
- *FSC-MOO₂*: FSC-MOO clustering technique optimizing PBM-index [15] and Sym-index as internal cluster validity indices and the number of features as another objective function.

The results obtained by two variants of FSC-MOO technique are reported in Table 4.

4.2.1 Results of sample selection and gene clustering

The number of time points selected by FSC-MOO clustering technique for Yeast Sporulation gene data is five out of seven. Based on these selected time points, FSC-MOO technique evolves six clusters ($K = 6$). The selected features are shown in Table 3. The S(C) index value for FSC-MOO

is 0.6212, which is highest to all nine clustering techniques. The Silhouette index values are reported in Table 4 for all the clustering techniques. The nine clustering techniques used here for the purpose of comparison utilize all the available time points during the clustering process. The number of obtained cluster and corresponding S(C) scores for these all clustering techniques MO-fuzzy, MOGA, FCM, Average Linkage, SOM, SGA, CRC, Spectral and K-mean clustering techniques are (6, 0.5877)(6, 0.5754)(7, 0.4696)(6, 0.5023)(6, 0.5794)(6, 0.5712)(8, 0.5623)

(6, 0.5606)(6, 0.4677). The DB index value for FSC-MOO is 0.2362, which is lowest to all corresponding nine clustering techniques and reported in Table 4. The DB index values corresponding to the partitionings obtained by corresponding clustering techniques are 0.2592, 0.2851, 0.4178, 0.3103, 0.3925, 0.2982, 0.3262, 0.3428 and 0.4456 respectively.

Apart from this, FSC-MOO₁ and FSC-MOO₂ are also executed on these data sets to obtain the optimal solutions. These variants select five and four features as the optimal number of features, respectively. Both these techniques

determine total six clusters ($K = 6$) in optimal partitioning on using these selected features. $S(C)$ and DB index values obtained by these two versions of FSC-MOO clustering approaches, FSC-MOO₁ and FSC-MOO₂ are [0.5944, 0.2441] and [0.6084, 0.2415] respectively. These values are reported in Table 4. $S(C)$ and DB index value for FSC-MOO is higher and lower than both of FSC-MOO₁ and FSC-MOO₂, which again proves the superiority of the FSC-MOO algorithm.

The number of time points selected by FSC-MOO clustering technique for Yeast Cell Cycle gene data is nine out of total seventeen time points. Based on these selected time points, FSC-MOO technique determines total five clusters ($K = 5$). The $S(C)$ index value for FSC-MOO is 0.4421, which is highest among all nine clustering techniques. The Silhouette index values and the optimal clusters (K) obtained by all the corresponding nine clustering techniques are reported in Table 4. Also, DB index value for FSC-MOO is the lowest compared to other nine techniques. It attains the DB index value of 0.2467. The DB index values corresponding to the partitionings obtained by MO-fuzzy, MOGA, FCM, SGA, Average Linkage, SOM, CRC, Spectral and K-mean clustering techniques are 0.2600, 0.2889, 0.4017, 0.2904, 0.5206, 0.4248, 0.2762, 0.4423 and 0.4232 respectively. All these values are shown in Table 4.

Apart from this, FSC-MOO₁ and FSC-MOO₂ are also executed on these data sets to obtain the optimal solutions. These variants select eleven and nine features as the optimal number of features, respectively. Using these selected features, both these techniques determine total six clusters ($K = 6$) in the optimal partitioning. $S(C)$ and DB index values obtained by these two versions of FSC-MOO clustering approaches, FSC-MOO₁ and FSC-MOO₂ are [0.4413, 0.2506] and [0.4171, 0.2924] respectively. These values are reported in Table 4. $S(C)$ and DB index value for FSC-MOO is higher and lower than both of FSC-MOO₁ and FSC-MOO₂, which again proves the superiority of the FSC-MOO algorithm.

The number of time points selected by FSC-MOO clustering technique for Arabidopsis Thaliana gene data is five out of total eight time points. Based on these selected time points, FSC-MOO technique evolves four clusters ($K = 4$). The $S(C)$ index value for FSC-MOO is 0.4258, which is highest among all nine clustering techniques. The Silhouette index values and the optimal clusters (K) obtained by all the corresponding nine clustering techniques are reported in Table 4. The obtained results on this data set again prove that with half number of selected features FSC-MOO performs better than the existing techniques. Also, DB index value of this partitioning is the lowest compared to other nine techniques. It attains the DB index value of 0.1469. The DB index values corresponding to the partitionings obtained by MO-fuzzy, MOGA, FCM, SGA, Average Linkage, SOM,

Table 5 Obtained Silhouette index values by different clustering techniques; here raw denotes original unprocessed data and pre denotes the pre-processed data

Data set_used	FSC-MOO		MO-fuzzy	
	K	S(c)	K	S(c)
Sporulation(raw)	6	0.5317	5	0.5135
Cell Cycle(raw)	5	0.4503	4	0.4091
Thaliana(raw)	4	0.5983	4	0.4968
D. melanogaster(raw)	3	0.8908	2	0.8299
C. elegans(raw)	3	0.8897	2	0.8645
D. melanogaster(pre)	3	0.3624	3	0.3238
C. elegans(pre)	3	0.8677	3	0.8045

Bold symbol indicates highest index value and optimal number of cluster obtained by FSC-MOO and Mo-fuzzy

CRC, Spectral and K-mean clustering techniques are 0.1802, 0.2678, 0.4234, 0.3982, 0.5125, 0.6246, 0.4146, 0.5755 and 0.4168, respectively. These all values are shown in Table 4.

Apart from this, FSC-MOO₁ and FSC-MOO₂ are also executed on these data set to obtain the optimal solutions. Both of these variants select six features for evolving optimal clusters. Based on these selected features, both these techniques give four clusters ($K = 4$). The $S(C)$ index value for these variants is given in Table 4. $S(C)$ values obtained by these two versions of FSC-MOO clustering approaches, FSC-MOO₁ and FSC-MOO₂ are 0.3958 and 0.4204, respectively. These values are reported in Table 4. But $S(C)$ value obtained by FSC-MOO algorithm is higher than those obtained by FSC-MOO₁ and FSC-MOO₂ techniques, which again proves the superiority of the FSC-MOO algorithm. Again DB index values are calculated for the partitioning results obtained by these two versions of FSC-MOO techniques. The obtained DB index values by FSC-MOO₁ and FSC-MOO₂ techniques are 0.2204 and 0.1692, respectively. These values are reported in Table 4. But DB index value obtained by FSC-MOO algorithm is lower than those obtained by FSC-MOO₁ and FSC-MOO₂ clustering techniques, which again proves the superiority of FSC-MOO algorithm.

Also, FSC-MOO and MO-fuzzy algorithms are applied on the Yeast Sporulation, Yeast Cell Cycle, and Arabidopsis Thaliana data without using any preprocessing steps. The selected features and the optimal number of clusters obtained by FSC-MOO algorithm is shown in Table 3. The Silhouette index values of the obtained partitioning results are shown in Table 5. Thus FSC-MOO determines the most reasonable partitioning from this data set where MO-fuzzy fails to evolve the true partitioning results. The corresponding Silhouette index value is also better than that obtained by MO-fuzzy clustering technique. Also, Silhouette index value for preprocessed data is higher than un-preprocessed data.

Table 6 Silhouette index and DB index values obtained by different clustering algorithms for pre-processed D. melanogaster and C. elegans data sets when applied for gene-clustering

Algorithm	D. melanogaster			C. elegans		
	K	S(c)	DB	K	S(c)	DB
FSC-MOO	3	0.9081	0.0653	4	0.9654	0.0143
MO-fuzzy	3	0.8609	0.1456	6	0.9245	0.2012
FCM	3	0.8565	0.2062	6	0.9089	0.2492
Average Linkage	3	0.8962	0.1034	6	0.9021	0.2624

Bold symbol indicates high index value and optimal number of clusters obtained by proposed algorithm

For D. melanogaster data set, we have applied FSC-MOO and MO-fuzzy clustering techniques on both the versions of the dataset obtained using with and without application of any preprocessing step. FSC-MOO for raw data D. melanogaster provides the optimal partitioning result with $K=3$, number of clusters. Silhouette index value for the obtained optimal solution is 0.8908. MO-fuzzy provides $K=2$ as the optimal number of clusters, and the corresponding Silhouette index value is 0.8299. These results are reported in Table 5. FSC-MOO technique selects 12 time points out of total thirty time points. The selected time points are shown in Table 3. Results again reveal that FSC-MOO technique attains the improved performance with half the number of time points.

Now, for preprocessed D. melanogaster data, we again apply the clustering techniques, FSC-MOO, MO-fuzzy, FCM, and Average Linkage. FSC-MOO provides partitioning results with $K=3$ number of clusters. The corresponding Silhouette index value is 0.9081. MO-fuzzy automatically determines $K=3$ as the number of clusters, and the corresponding Silhouette index value is 0.8609. FCM is executed with $K=3$ number of clusters, and the corresponding Silhouette index value is 0.8565. Average Linkage is executed with $K=3$ number of clusters, and the corresponding Silhouette index value is 0.8962. Results are reported in Table 6. FSC-MOO selects twelve time points out of thirty time points. The set of selected features is also shown in Table 3.

For C. elegans data set, we have applied FSC-MOO and MO-fuzzy clustering techniques on this data set with and without application of any preprocessing step. FSC-MOO for raw data C. elegans₁ provides partitioning results with $K=3$ number of clusters. Silhouette index value of the obtained optimal solution is 0.8897. MO-fuzzy provides $K=2$ as the optimal number of clusters, and the corresponding Silhouette index value is 0.8645. Results are reported in Table 5. FSC-MOO technique selects thirteen time points out of total thirty five time points. The selected features are shown in Table 3. Now, for preprocessed C. elegans₁ data, we have again applied FSC-MOO, MO-fuzzy, FCM, and Average Linkage clustering techniques. FSC-MOO attains $K=7$ number of clusters for the optimal partitioning, and the corresponding Silhouette index value is 0.9654. MO-fuzzy attains $K=6$ number of clusters on the optimal partitioning,

and the corresponding Silhouette index value is 0.9245. FCM and Average Linkage clustering techniques are also executed on this data set with $K=6$ as the number of clusters, and the corresponding Silhouette index values are 0.9089 and 0.9021, respectively. Results are reported in Table 6. FSC-MOO selects thirteen time points out of total thirty five time points. The feature combination selected by the proposed technique is shown in Table 3.

4.2.2 Results of gene selection and sample clustering

To show the effectiveness of the proposed FSC-MOO technique for gene selection and sample clustering, two gene expression data sets *SRBCT* and *Leukemia* are used. Here samples are treated as rows, and genes are treated as columns of the data set. The aim is to select the relevant set of genes which can cluster the samples well. FSC-MOO is applied in the same setting, as mentioned earlier. FSC-MOO automatically determines $K = 4$ number of clusters from this data set. Out of total 200 genes, FSC-MOO technique selects 86 genes on the optimal solution. Thus 200 genes are reduced to 86 genes. Therefore each sample is treated as a 86 dimensional vector. Here again, the values of an external cluster validity index, ARI-index, and percentage classification accuracy are calculated to validate the obtained partitioning results. The obtained ARI index value, %CA value and *Silhouette index* scores are 0.8171, 90.21 and 0.4331, respectively, for this data set (refer to Table 7). For the purpose of comparison, six clustering techniques, namely *MO-fuzzy*, *FCM*, *Average Linkage*, *SGA*, *Spectral* and *K-mean* clustering technique, utilizing all the 200 genes are used to cluster the available samples with $K = 4$ number of clusters. The ARI index, %CA and *Silhouette index* values obtained by these algorithms are (0.7677, 88.71, 0.4056), (0.4603, 0.7488, 0.3339), (0.1021, 49.05, 0.2008) and (0.3198, 70.81, 0.2957), (0.2132, 51.23, 0.2414) and (0.3135, 70.19, 0.2889) respectively. But all these values are less than the values obtained by our proposed FSC-MOO technique. The obtained results are reported in Table 7.

Again, FSC-MOO₁ and FSC-MOO₂ techniques are also applied on this particular data set (results reported in Table 7). They select 96 and 91 genes, respectively. In all the cases, the final partitioning result is associated with $K = 4$

Table 7 ARI index, %CA, Silhouette index, DB-index values obtained by different clustering algorithms for SRBCT and Leukemia data sets when applied for sample-clustering

Algorithm_used	SRBCT					Leukemia				
	K	ARI	%CA	S(C)	DB	K	ARI	%CA	S(C)	DB
FSC-MOO	4	0.8171	90.21	0.4331	0.1430	2	0.7248	89.60	0.4109	0.1760
FSC-MOO ₁	4	0.7402	85.27	0.4031	0.2521	2	0.6672	86.27	0.3749	0.2209
FSC-MOO ₂	4	0.7978	88.12	0.4107	0.2220	2	0.6908	87.81	0.3971	0.1994
MO-fuzzy	4	0.7677	86.71	0.4056	0.2482	2	0.6523	84.32	0.3436	0.2816
FCM	4	0.4603	79.21	0.3339	0.4213	2	0.5295	81.25	0.2796	0.3625
Average Linkage	4	0.1021	45.24	0.2008	0.7242	2	0.5971	82.82	0.3257	0.3016
SGA	4	0.3198	70.81	0.2957	0.6234	2	0.6021	83.06	0.3325	0.2692
Spectral	4	0.2132	51.23	0.2414	0.6934	2	0.5023	76.75	0.2634	0.4126
K-mean	4	0.3135	70.19	0.2889	0.6324	2	0.5523	78.25	0.2765	0.3985

Bold symbol indicates high index value and optimal number of cluster obtained by proposed algorithm

number of clusters. ARI index, %CA and *Silhouette index* values obtained by these two versions of FSC-MOO clustering approaches, FSC-MOO₁ and FSC-MOO₂ are (0.7402, 85.27, 0.4031) and (0.7978, 88.12, 0.4107), respectively. These values are reported in Table 7. But ARI and %CA values obtained by FSC-MOO algorithm are higher than those obtained by FSC-MOO₁ and FSC-MOO₂ techniques, which again proves the superiority of the FSC-MOO algorithm. Again DB index values are calculated for the obtained partitioning results by FSC-MOO, FSC-MOO₁, FSC-MOO₂, MO-fuzzy, FCM, Average Linkage, SGA, Spectral and K-mean clustering techniques. The obtained *DB index* values are 0.1430, 0.2521, 0.2220, 0.2482, 0.4213, 0.7242, 0.6234, 0.6934 and 0.6324 respectively. These values are reported in Table 7. But the *DB index* value obtained by FSC-MOO algorithm is lower than those obtained by FSC-MOO₁, FSC-MOO₂, MO-fuzzy, FCM, Average Linkage, SGA, Spectral and K-mean clustering techniques, which again proves the superiority of the FSC-MOO algorithm.

Again for Leukemia data set, FSC-MOO is applied as a sample-based-clustering to select the relevant set of genes which can partition the available samples into some relevant number of clusters. Here FSC-MOO automatically determines $K = 2$ number of clusters from this data set. Out of these 200 genes, FSC-MOO technique selects total 72 genes on the final optimal solution. Thus 200 genes are reduced to 72 genes, and therefore each sample is treated as a 72 dimensional vector. An external cluster validity index, ARI-index, and also percentage classification accuracy are calculated to validate the obtained partitioning results. The obtained ARI index and %CA values are 0.7248 and 89.60, respectively (refer to Table 7). To evaluate the performance another six clustering techniques, MO-fuzzy, FCM, Average Linkage and SGA, Spectral and K-mean clustering technique, are also applied on the same data set with $K = 2$ number of clusters with all the available number of genes for clustering purpose. The ARI index, %CA and Silhouette index values obtained by these algorithms are (0.6523, 84.32, 0.3436),

(0.5295, 81.25, 0.2796), (0.5971, 82.82, 0.3257), (0.6021, 83.06, 0.3325), (0.5023, 76.75, 0.2634) and (0.5523, 78.25, 0.2765), respectively. But all these values are less than the values obtained by executing the FSC-MOO technique. These values are reported in Table 7. Again, FSC-MOO₁ and FSC-MOO₂ techniques are also applied on this particular data set (results reported in Table 7). They select 84 and 77 genes, respectively. In all the cases, the final partitioning result is obtained with $K = 2$ number of clusters. ARI index, %CA and *Silhouette index* values obtained by these two versions of FSC-MOO clustering approaches, FSC-MOO₁ and FSC-MOO₂ are (0.6672, 86.27, 0.3749) and (0.6908, 87.81, 0.3971), respectively. These values are reported in Table 7. But ARI and %CA values obtained by FSC-MOO algorithm are higher than those obtained by FSC-MOO₁ and FSC-MOO₂ techniques, which again prove the superiority of FSC-MOO algorithm. Again DB index values are calculated for obtained partitioning results by FSC-MOO, FSC-MOO₁, FSC-MOO₂, MO-fuzzy, FCM, Average Linkage, SGA, Spectral and K-mean clustering techniques. The obtained DB index values are 0.1760, 0.2209, 0.1994, 0.2816, 0.3625, 0.3016, 0.2692, 0.4126 and 0.3985 respectively. These values are reported in Table 7. But the DB index value obtained by FSC-MOO algorithm is lower than those obtained by FSC-MOO₁, FSC-MOO₂, MO-fuzzy, FCM, Average Linkage, and SGA clustering techniques, which again proves the superiority of the FSC-MOO algorithm.

4.2.3 Biological significance

To show the biological relevance of an obtained cluster, Gene Ontology (GO) annotation database⁸ is used. So, in order to compute the compatibility between number of genes n for a particular GO category and cluster of size K , probability p is calculated as given in Equation [33].

⁸ <http://db.yeastgenome.org/cgi-bin/GO/goTermFinder>

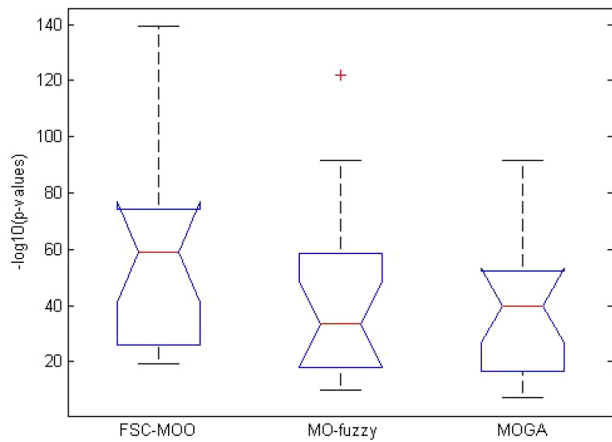


Fig. 5 Boxplots of p values of the most significant GO terms of all clusters obtained by FSC-MOO, MO-fuzzy and MOGA algorithms for Yeast Sporulation data when applied for gene-clustering

$$p = 1 - \sum_{j=0}^{n-1} \frac{\binom{t}{j} \binom{l-t}{K-j}}{\binom{l}{n}} \quad (8)$$

Where, t and l represent the number of genes for a particular GO category and the total number of genes to genome. After getting the p -value for each GO category, we perform biological significance test for genes in a cluster. In any situation, if p value equals zero then it signifies that genes within a particular cluster show the same biological function.

In this paper at 1% significance level, we have performed the biological significance test for Yeast Sporulation data set for gene-clustering. Also, the clustering results obtained by different algorithms are biologically tested. In the proposed FSC-MOO clustering technique all the obtained 6 gene clusters are more biological significant, while the number of biologically significant clusters for MO-fuzzy, MOGA, FCM, SGA, Average linkage, SOM, CRC, Spectral clustering and K-mean clustering are 6, 6, 4, 6, 4, 4, 6, 6 and 6 respectively.

Here, for Yeast Sporulation data, the number of GO terms for individual gene clusters varies; first cluster contains 61 terms, second cluster contains 50 terms, third cluster contains 51 terms, fourth cluster contains 57 terms, fifth one contains 19 terms, and the last one contains total 29 GO terms at false discovery rate (FDR) < 0.01. Now, for MO-fuzzy, the number of GO terms per cluster varies in comparison with FSC-MOO; first cluster contains 54 terms, second cluster contains 34 terms, third cluster contains 29

Table 8 Three most significant GO terms of individual six clusters of Yeast sporulation data and their p values obtained after application of FSC-MOO clustering technique when applied for gene-clustering

Clusters	Significance GO term	p value
Cluster1	Cytoplasmic translation:GO:0002181	3.36E-61
	Translation: GO:0006412	9.80E-32
	Cellular protein metabolic process: GO:0044267	2.07E-17
Cluster2	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
Cluster3	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
Cluster4	Ribosome biogenesis :GO:0042254	1.45E-12
	Ribonucleoprotein complex biogenesis :GO:0022613	5.44E-11
	rRNA processing:GO:0006364	4.22E-09
Cluster5	Meiotic nuclear division:GO:0007126	2.90E-26
	Meiotic cell cycle:GO:0051321	2.90E-26
	Reciprocal DNA recombination :GO:0035825	6.28E-26
Cluster6	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

Table 9 p values produced by Wilcoxon’s rank sum test comparing FSC-MOO with other algorithms for gene-clustering

Data set	MO-fuzzy	MOGA	FCM	SGA	SOM	CRC	Spectral	K-mean
<i>Sporulation</i>	3.69E-05	3.87E-05	2.12E-09	3.77E-04	4.12E-04	5.22E-04	6.21E-04	3.46E-08
<i>CellCycle</i>	1.07E-03	2.22E-03	5.60E-08	2.77E-04	4.44E-05	2.11E-03	5.23E-05	2.46E-03
<i>Arabidopsis</i>	1.01E-03	2.07E-03	7.22E-06	3.54E-03	6.12E-07	2.64E-03	1.08E-14	3.54E-07

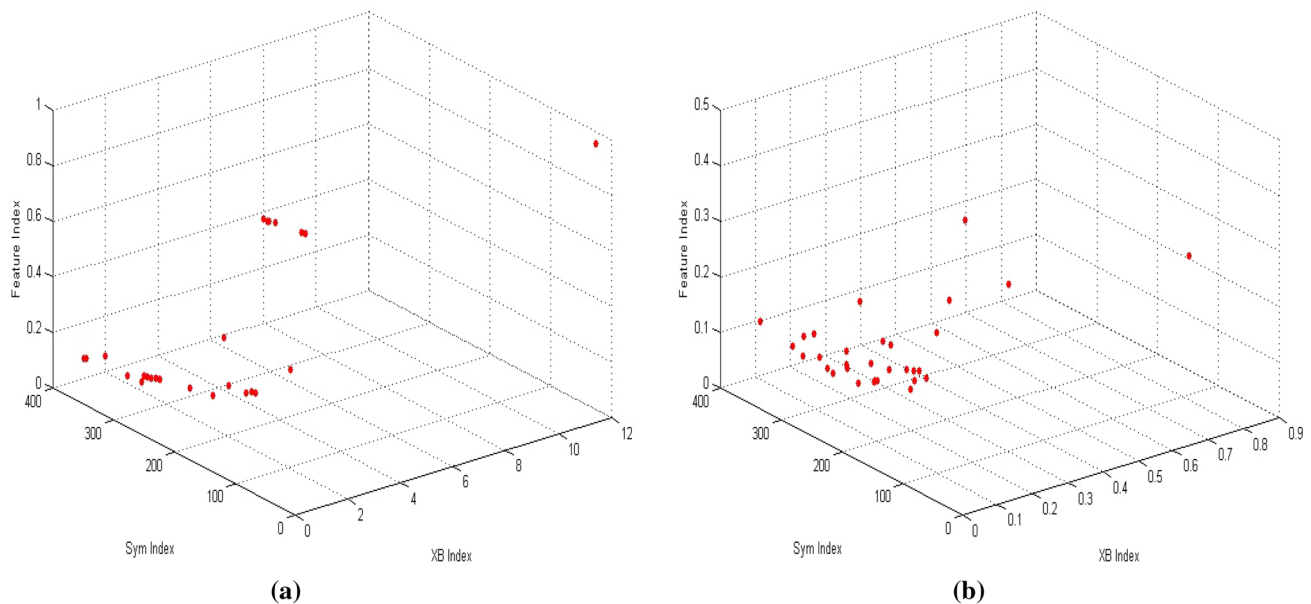


Fig. 6 Pareto optimal fronts obtained after application of FSC-MOO algorithm for **a** Yeast Sporulation, **b** Yeast Cell Cycle data sets for gene-clustering

terms, fourth cluster contains 22 terms, fifth cluster contains 9 terms, and the last cluster contains 50 GO terms at false discovery rate (FDR) < 0.01. The p -values for the three most relevant GO terms of genes belonging to a particular cluster are shown in Table 8. We consider that all p -values ≤ 0.01 . These results have been obtained after the application of the proposed FSC-MOO clustering technique on Yeast Sporulation data set. Lower the p value or higher $-\log_{10}(p$ value) ensures that the clusters obtained by different algorithms, consisting of significant GO terms, are more biologically relevant. The p values are log transformed for better readability. To show that the six gene-clusters obtained by FSC-MOO technique for Yeast Sporulation data are more biologically significant, box plot is shown for all clusters having significant GO terms that are associated with p values. These are compared with those obtained by MOGA and MO-fuzzy algorithms because both these techniques provide six clusters having significant GO terms. Box plots for all six clusters having significant GO terms and associated with p values, produced by FSC-MOO, MOGA and MO-fuzzy clustering techniques are shown in Fig. 5. From Fig. 5, it is clear that FSC-MOO gives higher $-\log_{10}(p$ value) value in comparison with MOGA and MO-fuzzy. So it has been proved that clusters obtained by our proposed FSC-MOO approach for Yeast Sporulation data are more biologically significant and functionally enriched. Apart from this, with the help of gene ontology database⁹, biologically significant

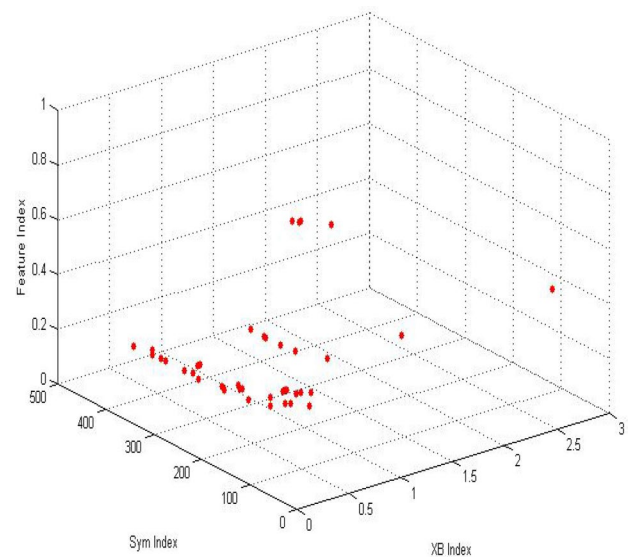


Fig. 7 Pareto optimal fronts obtained after application of FSC-MOO algorithm for Arabidopsis Thaliana data sets for gene-clustering

test for Rat CNS data is also performed. For the proposed FSC-MOO, all the obtained six clusters are biologically significant. Here again, p values ≤ 0.01 are considered for GO term evaluations. Here, for Rat CNS data, the number of GO terms for individual gene cluster varies; first cluster contains 12 terms, second cluster contains 42 terms, third cluster contains 43 terms, fourth cluster contains 6 terms, fifth one contains 19 terms and the last one contains total

⁹ <http://geneontology.org/>

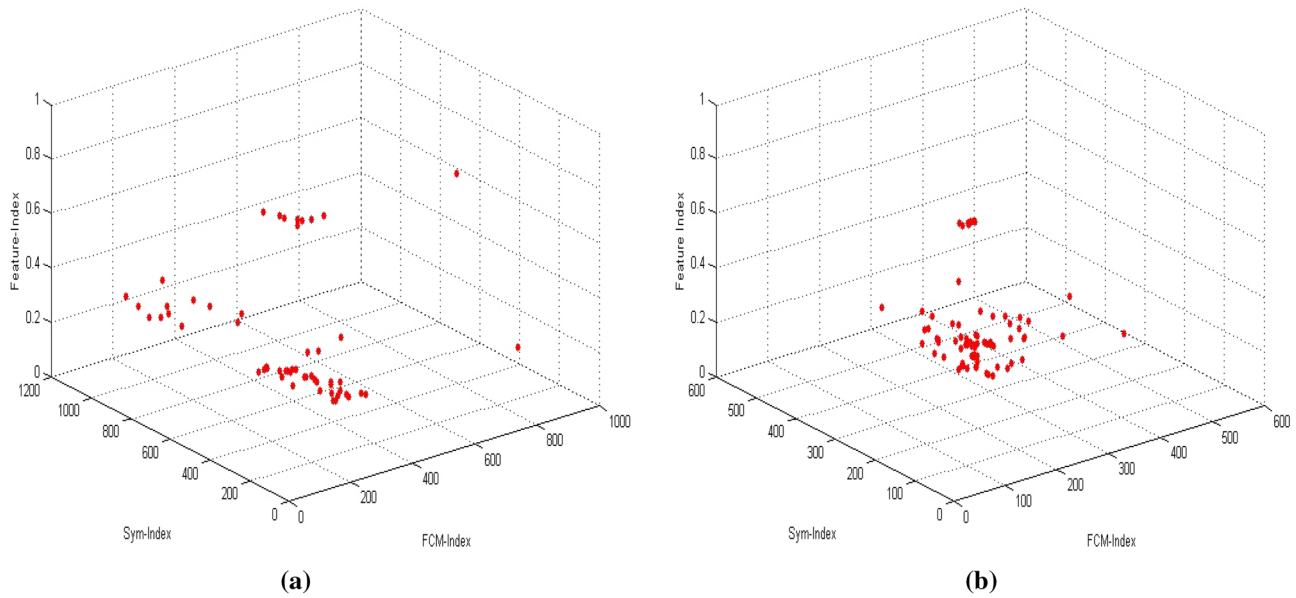


Fig. 8 Pareto optimal fronts obtained after application of FSC-MOO₁ algorithm for **a** Yeast Sporulation, **b** Arabidopsis Thaliana data sets for gene-clustering

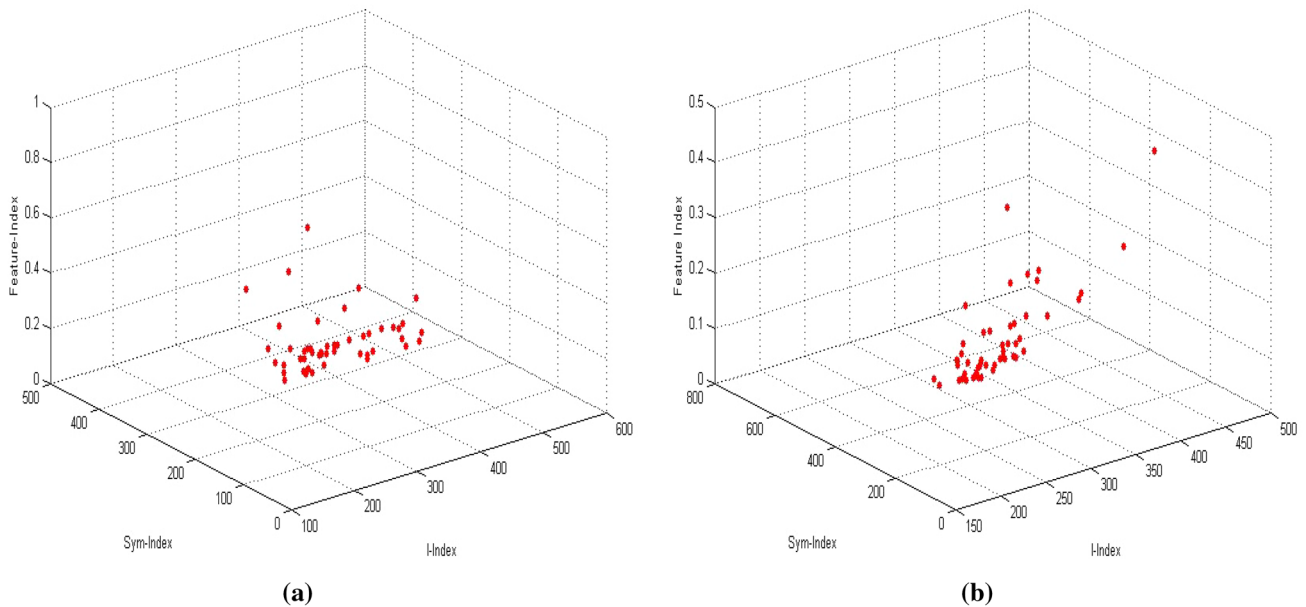


Fig. 9 Pareto optimal fronts obtained after application of FSC-MOO₂ algorithm for **a** Yeast Sporulation, **b** Yeast Cell Cycle data sets for gene-clustering

11 GO terms at false discovery rate (FDR) < 0.01. Now, for MO-fuzzy, the number of GO terms per cluster varies in comparison with FSC-MOO; first cluster contains 9 terms, second cluster contains 15 terms, third cluster contains 39 terms, fourth cluster contains 3 terms, fifth cluster contains

9 terms and the last cluster contains 15 GO terms at false discovery rate (FDR) < 0.01. The biologically significant test again reveals that the clusters obtained by the proposed FSC-MOO clustering technique after application on gene

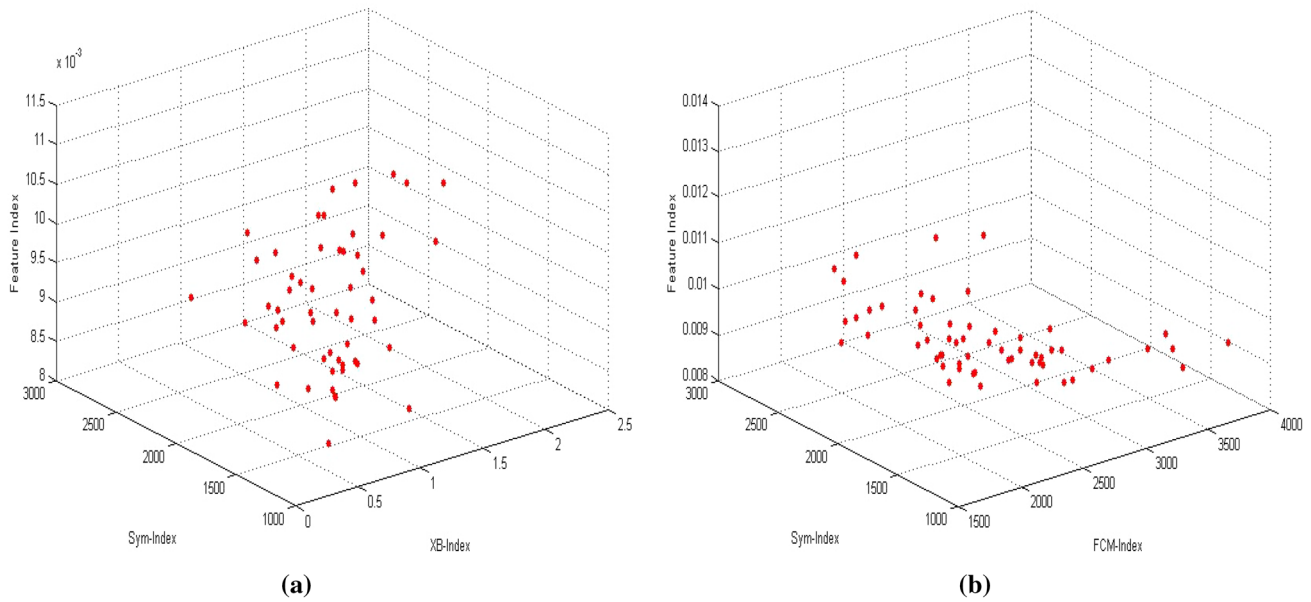


Fig. 10 Pareto optimal fronts obtained after application of FSC-MOO, FSC-MOO₁ on SRBCT data set for gene-clustering

expression data sets are more biologically relevant and functionally enriched.

4.2.4 Statistical significance

To show the efficacy of the proposed clustering technique FSC-MOO, again statistical significance test is performed using Wilcoxon's rank sum test [36]. In this regard, the comparative analysis of p values obtained by FSC-MOO with respect to different gene clustering techniques are highlighted in Table 9.

The p values reported in Table 9 are attainable at a level of significance less than 5%. From Table 9, we can see that obtained p values are more statistically significant because of the difference between the median values of two different groups. If differences show the resultant value other than zero, then it is totally against null hypothesis and is strongly accepted as more statistically significant. In this experiment, we have considered the median values of Silhouette index attained by FSC-MOO and other gene-clustering techniques. Based on the results of this test, we can easily conclude that Silhouette index values obtained by FSC-MOO are more statistically significant.

4.2.5 Visualizations of the obtained clustering results

For three preprocessed gene expression data sets, Yeast Sporulation, Yeast Cell Cycle and Thaliana, the Pareto

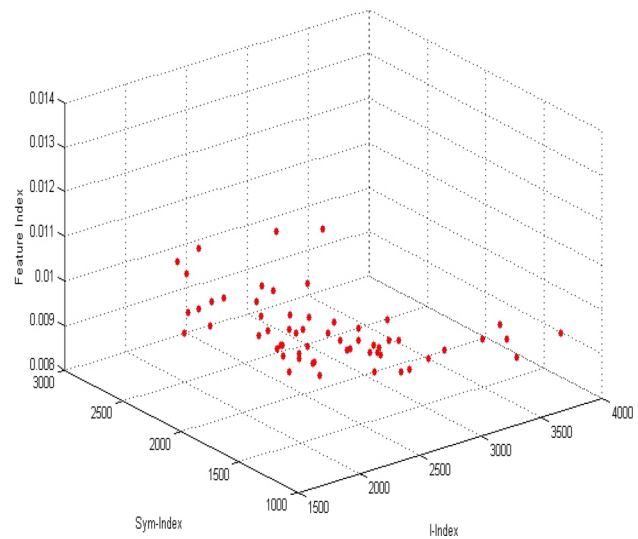


Fig. 11 Pareto optimal fronts obtained after application FSC-MOO₂ on SRBCT data set for gene-clustering

optimal fronts obtained by FSC-MOO, FSC-MOO₁ and FSC-MOO₂ clustering techniques are shown in Figs. 6, 7 8 and 9, respectively. For SRBCT data set, Pareto optimal fronts obtained by FSC-MOO, FSC-MOO₁ and FSC-MOO₂ clustering techniques are shown in Figs. 10a, b and 11, respectively. Also, Silhouette index values of Pareto optimal fronts obtained after application of FSC-MOO, FSC-MOO₁,

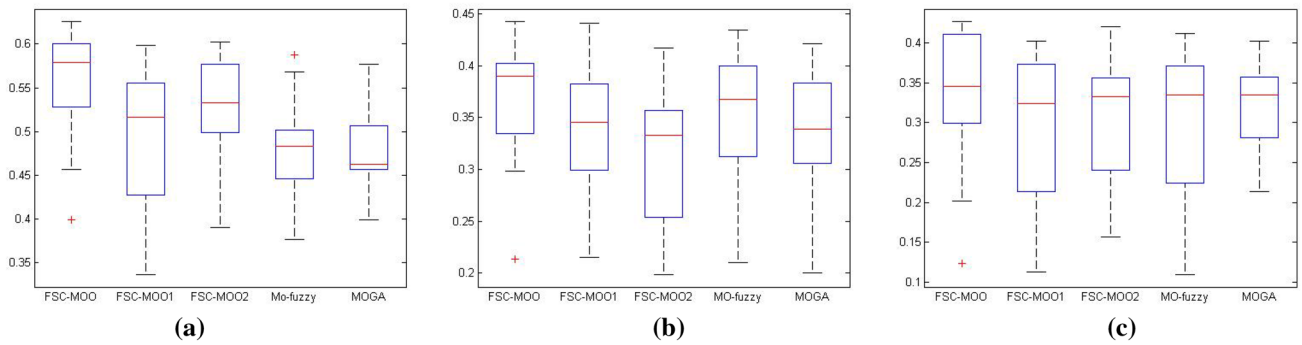


Fig. 12 Silhouette index values of Pareto optimal fronts obtained after application of FSC-MOO, FSC-MOO₁, FSC-MOO₂, MO-fuzzy and MOGA algorithm for **a** Yeast Sporulation, **b** Yeast Cell Cycle (c) Arabidopsis Thaliana data sets for gene-clustering

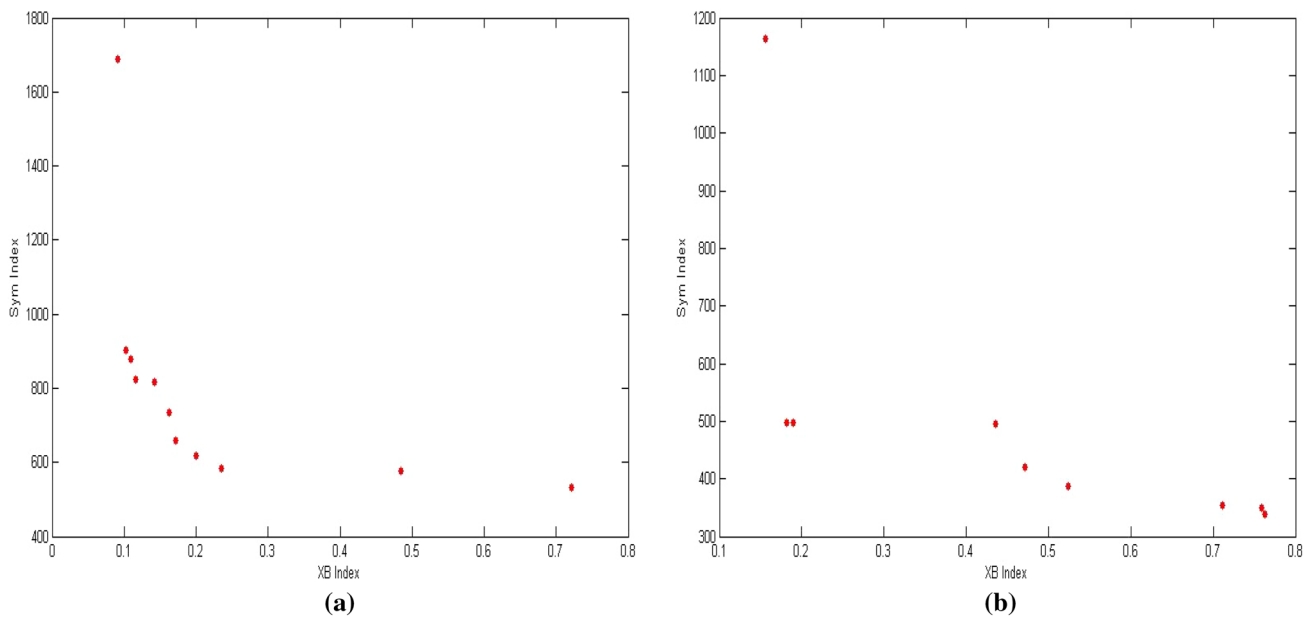


Fig. 13 Pareto optimal fronts obtained after application of MO-fuzzy algorithm for un-preprocessed **a** Yeast Sporulation, **b** Yeast Cell Cycle data sets

FSC-MOO₂, MO-fuzzy and MOGA on Yeast Sporulation, Yeast Cell Cycle and Thaliana gene expression data sets are shown in Fig. 12. Also, the Pareto optimal fronts obtained for un-preprocessed gene expression data sets by MO-fuzzy and FSC-MOO algorithms are shown in Figs. 13, 14, 15 and 16.

Also, Eisen and cluster profile plots are used to visualize the obtained partitioning results by the proposed technique

FSC-MOO. In Eisen plot, patterns having similar colors are grouped, have similar functionality. In another way, we can say that genes within a particular cluster perform similar functionality because their gene expression values have similar colors. Eisen plots are shown in Fig. 17 for obtained partitionings for three preprocessed gene expression data sets (Yeast Sporulation, Yeast Cell Cycle and Arabidopsis). Similarly, the cluster profile plot is a representation

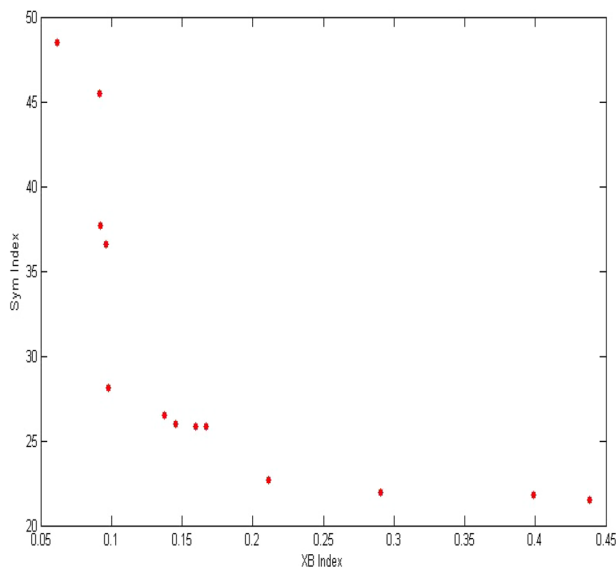


Fig. 14 Pareto optimal fronts obtained after application of MO-fuzzy algorithm for un-preprocessed Arabidopsis Thaliana data sets

of normalized gene expression values over different time points. Cluster profile plots are shown in Figs. 18, 19 and 20 for obtained partitionings, respectively.

4.3 Observations

Results of simultaneous sample selection and gene clustering on seven benchmark gene expression data sets show that, the proposed simultaneous feature selection and automatic clustering technique performs better as compared to the other nine existing gene expression data clustering techniques. As the proposed technique automatically selects the number of samples/time points required to cluster the given gene expression data, the time complexity of the entire clustering process is reduced further. Note that the partitioning results of the proposed FSC-MOO clustering are better than those obtained by two recently developed multiobjective based gene expression data clustering techniques, MO-fuzzy [29], MOGA [3] (as shown in Tables 4 and 6). MO-fuzzy technique also utilizes AMOSA as the search technique. Here point symmetry-based distance [1] is used for assigning points to different clusters. But no feature selection step was executed during the clustering process. All the given features were used for distance computations. It utilizes two internal cluster validity indices, an Euclidean distance based cluster validity index, XB-index and a symmetry distance based cluster validity index, Sym-index. MOGA uses a multiobjective evolutionary algorithm, NSGA-II [8], as the underlying optimization technique. Here Euclidean distance is used for cluster assignment. Center-based encoding is used, and two cluster validity indices are used for optimization. These are XB-index and FCM-index. Here also, no feature selection step was executed during the clustering process. Thus the

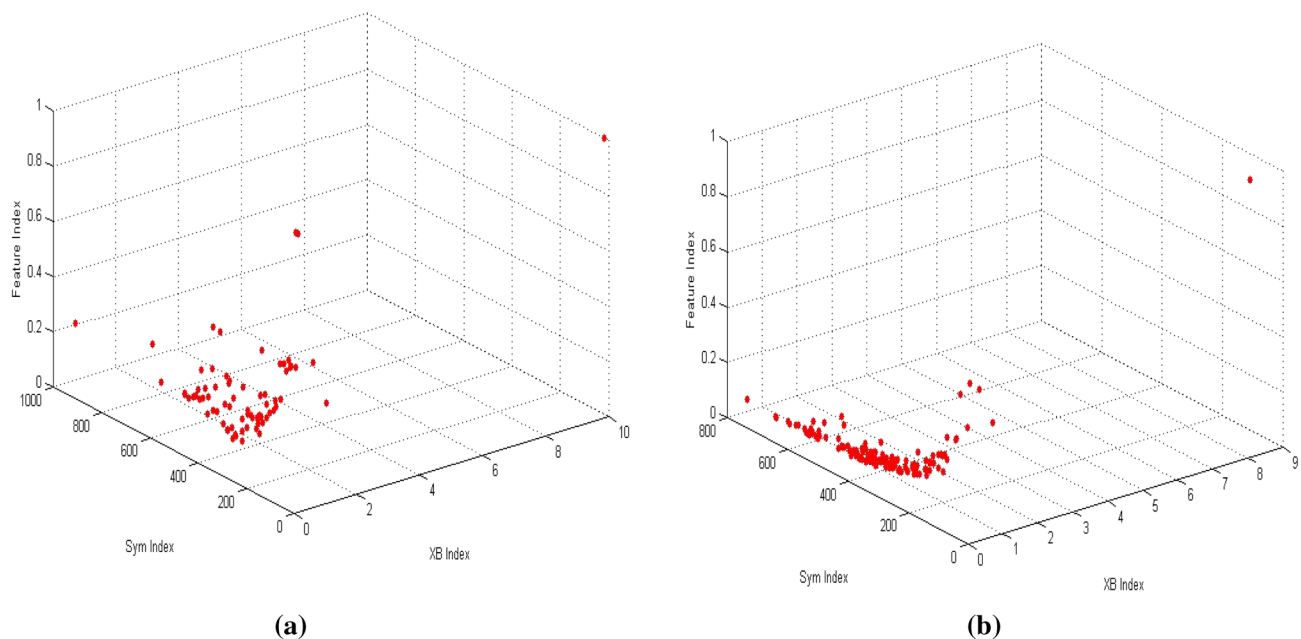


Fig. 15 Pareto optimal fronts obtained after application of FSC-MOO algorithm for un-preprocessed **a** Yeast Sporulation, **b** Yeast Cell Cycle

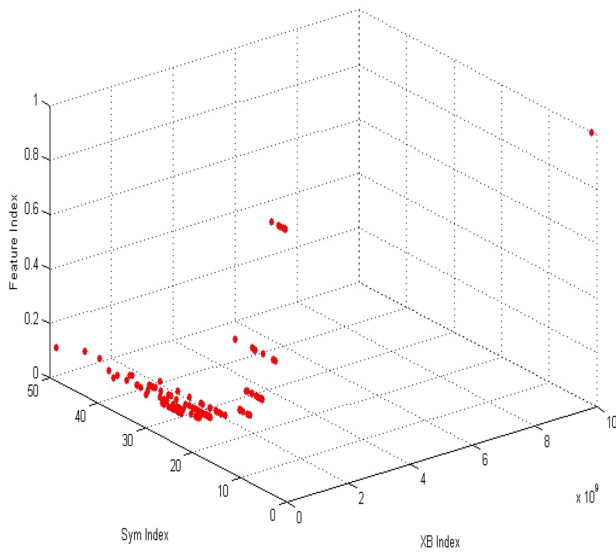


Fig. 16 Pareto optimal fronts obtained after application of FSC-MOO algorithm for un-processed Arabidopsis Thaliana data sets

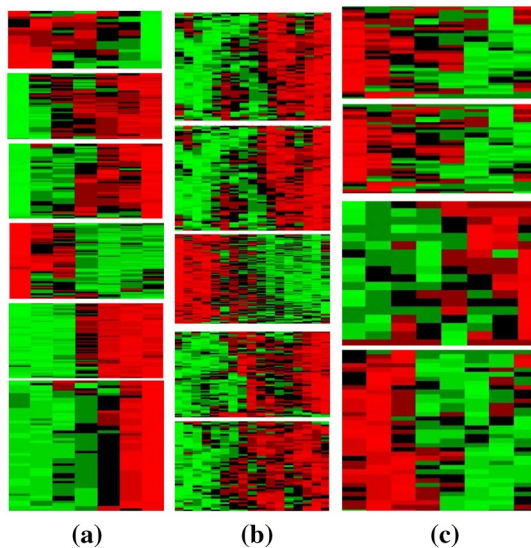


Fig. 17 Eisen Plot for **a** Yeast Sporulation, **b** Yeast Cell Cycle, **c** Arabidopsis Thaliana after application of FSC-MOO clustering technique for gene-clustering

improved results by the proposed FSC-MOO technique further illustrate the usefulness of the feature selection step. In current work, we have combined feature selection and clustering in a single step. We have shown with sufficient experimental results that the feature selection is required for solving the gene expression data clustering problem effectively. Existing MOO based techniques, when applied with

all the available features on the benchmark gene expression data sets, attain some decreased values of Silhouette index as compared to our proposed FSC-MOO technique. These results are due to the use of all features for distance computations in MOGA and MO-fuzzy clustering techniques. We have also applied the FSC-MOO clustering technique on two new gene expression data sets obtained using RNA sequencing, *D. melanogaster* and *C. elegans*. Here in this paper, we have used two different versions of all the data sets, one with some pre-processing step and another without using any pre-processing step. The proposed technique is applied on the above mentioned data sets. Results are then compared with nine other clustering techniques for gene expression data sets. The comparison of results is done using one internal cluster validity index, Silhouette index. The obtained results prove that this proposed clustering technique attains improved performance with less number of features. From the results mentioned above, we can further prove that the feature selection, as well as pre-processing steps are very crucial for clustering all the gene expression data sets.

The better results obtained by FSC-MOO technique thus show the utility of feature selection while clustering the gene expression data. The improved performance of the proposed technique is due to the following reasons:

- Here, a set of features is first selected, then clustering is applied on the given gene expression data based on those features. This helps to reduce the time complexity of the entire clustering process.
- All the features (genes or samples) present in the gene expression data may not be relevant for clustering. So, the selection of the appropriate subset of features helps to get improved partitioning.
- The proposed technique is able to identify automatically (i) appropriate set of cluster centers (ii) appropriate feature combinations and the (iii) appropriate partitioning.
- Use of point symmetry-based distance for assignment of genes to different clusters helps to determine clusters having different shapes.
- Use of multiple cluster quality measures as the objective functions helps to determine the appropriate partitioning automatically.

5 Conclusions

For clustering purposes, not all the features (it may be samples or genes depending on gene-clustering or sample-clustering) are essential. So in this paper, simultaneous feature selection and unsupervised clustering is posed as a multiobjective

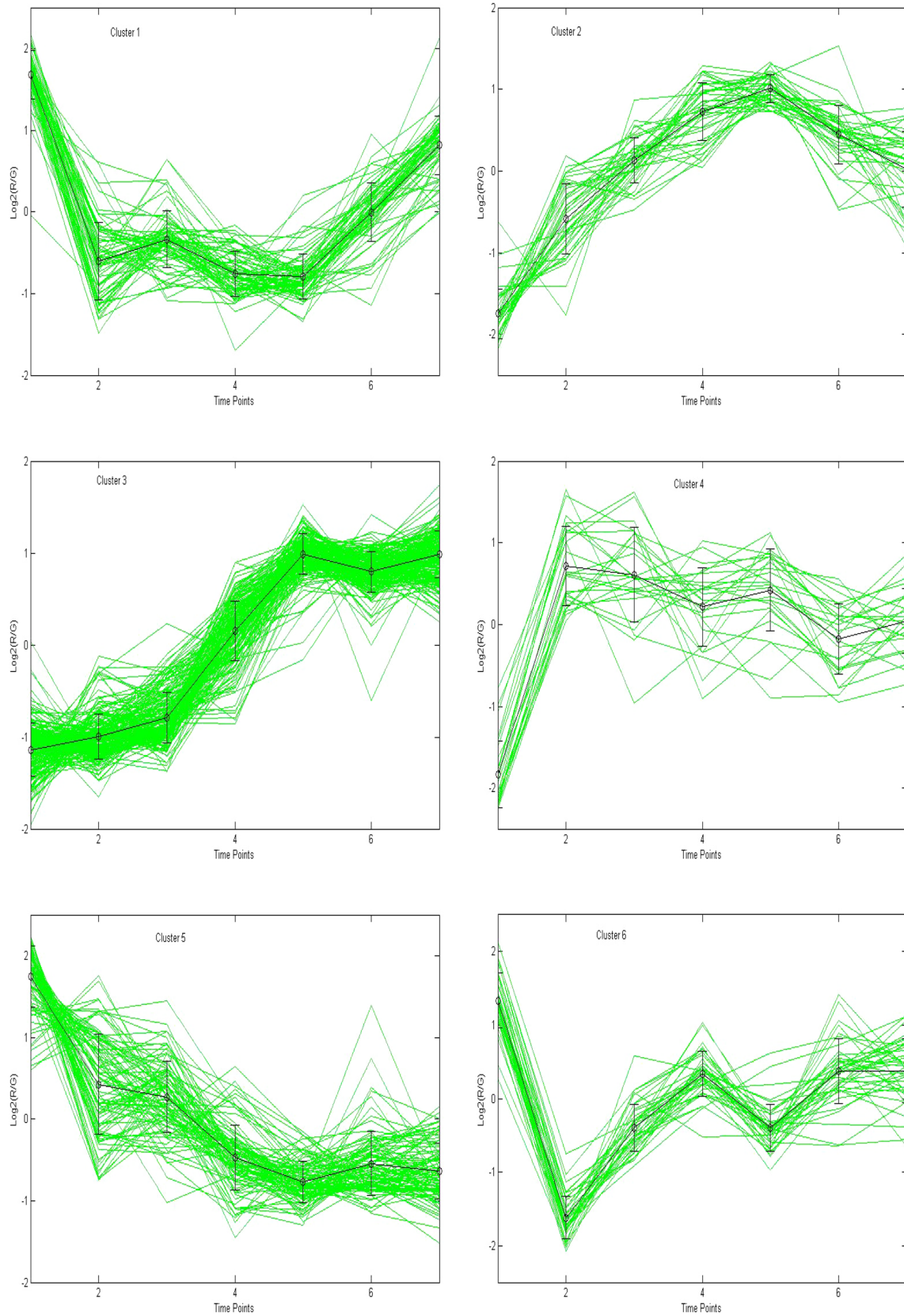


Fig. 18 Cluster profile plot for Yeast Sporulation data obtained after application of FSC-MOO clustering technique

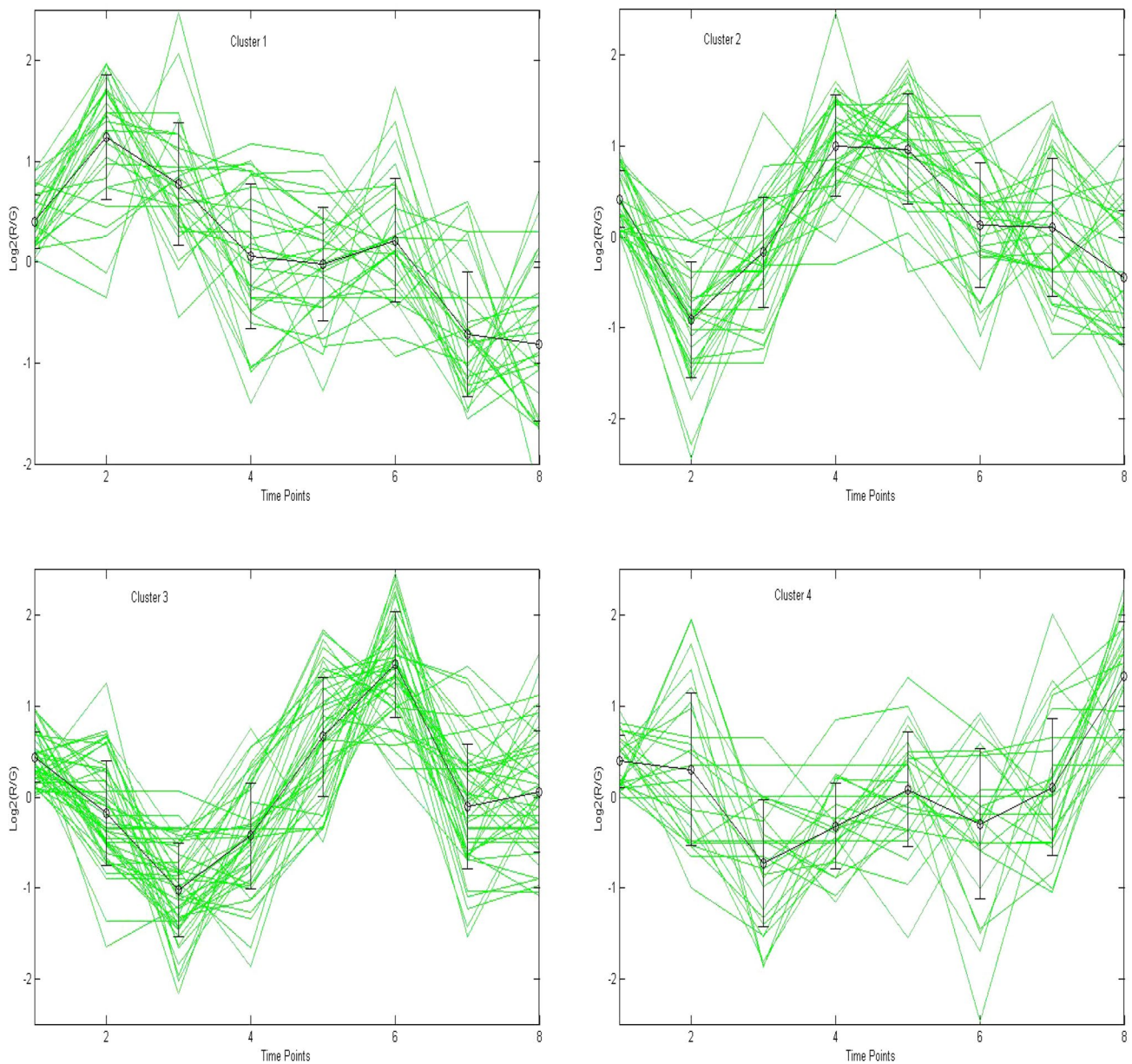


Fig. 19 Cluster profile plot of Arabidopsis Thaliana data obtained after application of FSC-MOO clustering method

optimization problem. Here in proposed multiobjective based approach FSC-MOO, some relevant time points of gene expression data are selected from a large collection of time points, and thereafter clustering is performed in the reduced space to obtain a good partitioning result. Three objective functions Sym-index, XB-index and feature index are optimized simultaneously to obtain a large collection of non-dominated solutions. The proposed technique FSC-MOO is able to select the relevant features and evolve the true optimal clusters. The effectiveness of the proposed algorithm, namely

FSC-MOO is demonstrated for simultaneous sample selection, and gene clustering on three open access gene expression data sets obtained from micro-array technology and four bench-mark gene expression data sets obtained from RNA-sequencing. Again, the results obtained by FSC-MOO technique are compared with some existing clustering techniques like MOGA, MO-fuzzy, FCM, SGA, Self Organizing Map (SOM), Chinese Restaurant Clustering (CRC), hierarchical average linkage, Spectral and K-mean clustering techniques. Also, biological significant test is carried out to show that the

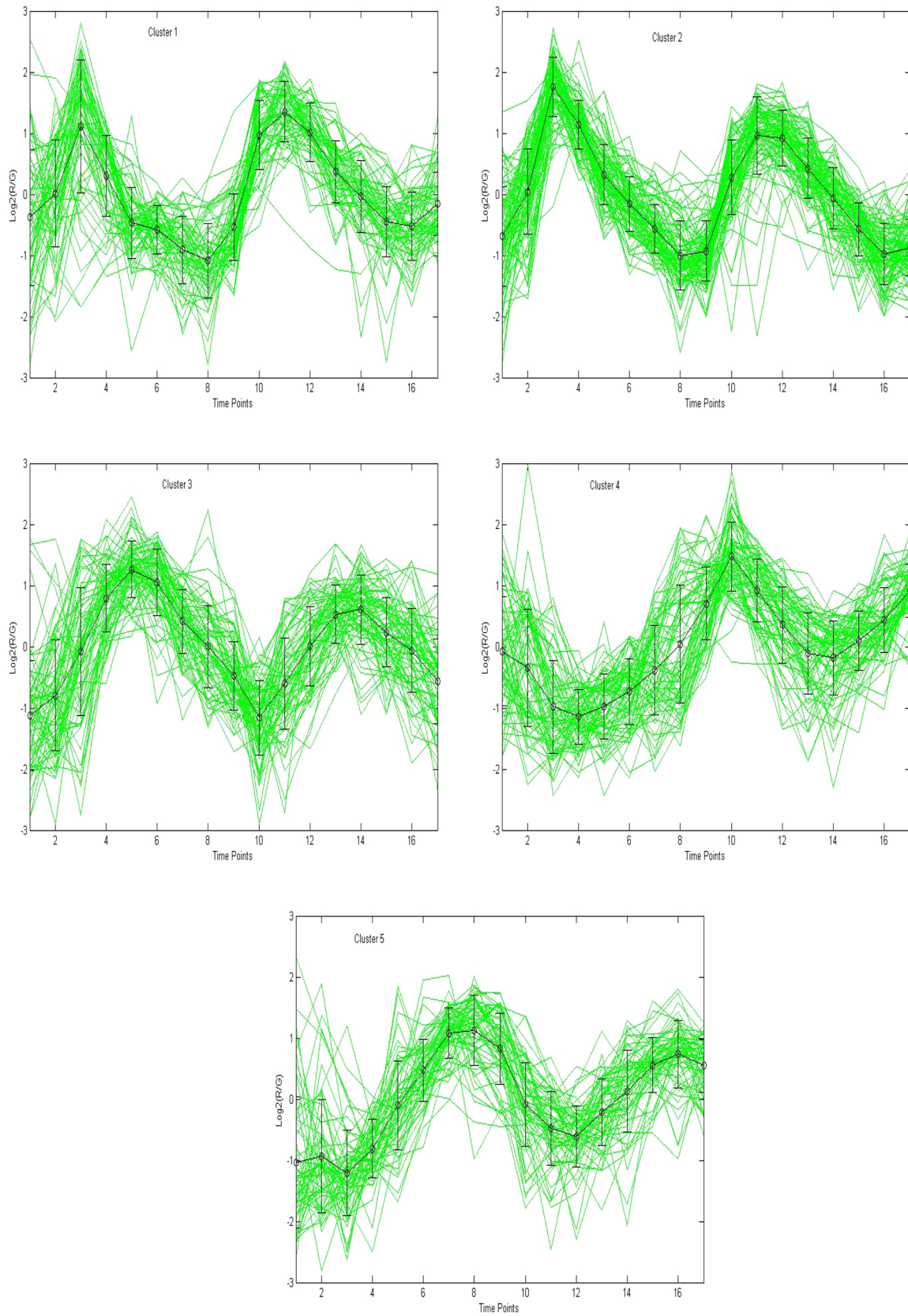


Fig. 20 Cluster profile plot of Yeast Cell Cycle data obtained after application of FSC-MOO clustering technique

obtained clusters are more biologically relevant and functionally enriched. In another part of the experimental results, the effectiveness of the proposed algorithm for simultaneous gene selection and sample clustering is shown on two gene expression data sets.

Future work may include the use of some supervised information while clustering the gene expression data sets. We would also like to apply the proposed technique for some cancer data sets or mRNA data sets.

Acknowledgements Dr. Sriparna Saha would like to acknowledge the support of SERB Women in Excellence Award-SB/WEA-08/2017 for conducting this research.

References

- Bandyopadhyay S, Saha S (2007) Gaps: A clustering method using a new point symmetry-based distance measure. *Pattern Recognit* 40(12):3430–3451
- Bandyopadhyay S, Saha S (2008) A point symmetry-based clustering technique for automatic evolution of clusters. *Knowl Data Eng IEEE Trans* 20(11):1441–1457
- Bandyopadhyay S, Mukhopadhyay A, Maulik U (2007) An improved algorithm for clustering gene expression data. *Bioinformatics* 23(21):2859–2865
- Bandyopadhyay S, Saha S, Maulik U, Deb K (2008) A simulated annealing-based multiobjective optimization algorithm: Amosa. *Evolut Comput IEEE Trans* 12(3):269–283
- Bezdek JC (2013) *Pattern recognition with fuzzy objective function algorithms*. Springer, Berlin
- Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. *Science* 282(5389):699–705
- Davies DL, Bouldin DW (1979) A cluster separation measure. *Patt Anal Mach Intell IEEE Trans* 2:224–227
- Deb K (2001) *Multi-objective optimization using evolutionary algorithms*, vol 16. Wiley, New York
- Handl J, Knowles J (2006) Feature subset selection in unsupervised learning via multiobjective optimization. *Int J Comput Intell Res* 2(3):217–238
- Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218
- Huynen M, Snel B, Lathe W, Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* 10(8):1204–1210
- Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors *sfb* and *mbf*. *Nature* 409(6819):533–538
- Li JJ, Huang H, Bickel PJ, Brenner SE (2014) Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modencode rna-seq data. *Genome Res* 24(7):1086–1101
- MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA vol 1, pp 281–297
- Maulik U, Bandyopadhyay S (2002) Performance evaluation of some clustering algorithms and validity indices. *Pattern Anal Mach Intell IEEE Trans* 24(12):1650–1654
- Maulik U, Bandyopadhyay S (2003) Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification. *Geosci Remote Sens IEEE Trans* 41(5):1075–1081
- Maulik U, Mukhopadhyay A, Bandyopadhyay S (2009) Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes. *BMC Bioinf* 10(1):27
- McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE, Engelhardt BE (2018) Clustering gene expression time series data using an infinite gaussian process mixture model. *PLoS Comput Biol* 14(1):e1005896
- Mitra S, Saha S (2019) A multiobjective multi-view cluster ensemble technique: application in patient subclassification. *PLoS ONE* 14(5):e0216904
- Mukhopadhyay A, Bandyopadhyay S, Maulik U (2010) Multi-class clustering of cancer subtypes through SVM based ensemble of pareto-optimal solutions for gene marker identification. *PLoS ONE* 5(11):e13803
- Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghien E, Ameh F, Achas M, Adebisi E (2016) Clustering algorithms: their application to gene expression data. *Bioinform Biolo Insights* 10:BBI-S38316
- Parraga-Alava J, Dorn M, Inostroza-Ponta M (2018) A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies. *BioData Min* 11(1):16
- Pati SK, Das AK (2012) Optimal samples selection from gene expression microarray data using relational algebra and clustering technique. In: *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012)* held in Visakhapatnam, India, January 2012. Springer, pp 507–514
- Paul S, Maji P (2014) City block distance and rough-fuzzy clustering for identification of co-expressed micrnas. *Mol BioSyst* 10(6):1509–1523
- Paul S, Vera J (2015) Rough hypercuboid based supervised clustering of mirnas. *Mol BioSyst* 11(7):2068–2081
- Qin ZS (2006) Clustering microarray gene expression data using weighted chinese restaurant process. *Bioinformatics* 22(16):1988–1997
- Reymond P, Weber H, Damond M, Farmer EE (2000) Differential gene expression in response to mechanical wounding and insect feeding in arabidopsis. *Plant Cell* 12(5):707–719
- Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20:53–65
- Saha S, Ekbal A, Gupta K, Bandyopadhyay S (2013) Gene expression data clustering using a multiobjective symmetry based clustering technique. *Comput Biol Med* 43(11):1965–1977
- Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci* 96(6):2907–2912
- Tang C, Zhang A (2002) An iterative strategy for pattern discovery in high-dimensional data sets. In: *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, pp 10–17
- Tang C, Zhang L, Zhang A, Ramanathan M (2001) Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In: *Proceedings of the IEEE 2nd international symposium on bioinformatics and bioengineering conference, 2001*. IEEE, pp 41–48
- Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* 22(3):281–285
- Tou JT, Gonzalez RC (1974) *Pattern recognition principles*. Addison-Wesley, Reading
- Von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17(4):395–416

36. Wilcoxon F, Katti S, Wilcoxon RA (1963) Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. American Cyanamid, Pearl River, NY
37. Xie XL, Beni G (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 8:841–847

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.