Supplementary file: Simultaneous Clustering and Feature Weighting using Multiobjective Optimization for Identifying Functionally Similar miRNAs

Sriparna Saha, *Senior Member, IEEE,*, Sudipta Acharya, Kavya K, and Saisree

I. Description of chosen datasets

Please note that we have considered microarray gene expression data which contains numeric features (downloaded from www.ncbi.nlm.nih.gov/geo/). Our proposed method *AutoMODE* can be successfully applied on any other real life datasets where features are numeric. However, there are several real life categorical attribute based biological datasets available in UCI Machine Learning Repository [1]. Our proposed clustering algorithm can work on them too with some preprocessing. For any dataset, containing categorical attributes can be converted first to numeric values on which our proposed method can be applied successfully.

A. GSE16473

It is the analysis to evaluate the role of miRNAs in skeletal muscle regeneration. Hence, global miRNA expression is measured during muscle cell growth and differentiation. This data set contains 231 miRNAs and 7 time points.

B. GSE17155

It is the analysis to test the hypothesis that there is a specific miRNA expression signature which characterizes male breast cancers. The miRNA microarray analysis was performed in a series of male breast cancers and compared them to cases of male gynecomastia and female breast cancers.This data set contains 774 miRNAs and 38 time points.

C. GSE29495

The miRNA profiling of kidney tissue from C57BL/6 mice that received a 30 minute is chemic injury compared with control kidney tissue from mice that received sham operation only has been conducted. The number of miRNAs and the time points are 574 and 17, respectively.

II. Chosen Cluster Validity Measures

Those are described as follows,

All authors are from the department of Computer Science and engineering, IIT Patna, India, Corresponding author: sudiptaacharya.2012@gmail.com

[1]http://www.ics.uci.edu/ mlearn/MLRepository.html

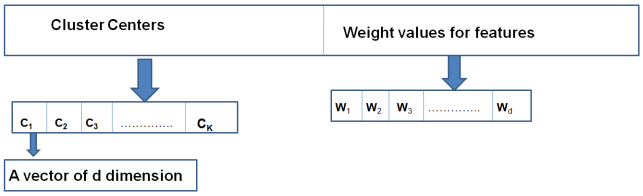

Fig. 1: Chromosome Representation of Proposed Differential Evolution Based Multiobjective Clustering Technique (*AutoMODE*) for miRNA Classification

A. Silhouette index

Silhouette index [1] is an internal cluster validity index which we have used in this article to measure the goodness of clustering solutions. It is a function of cluster compactness and separation. Suppose,
x= average distance of a particular point from other points of the same cluster in which that point resides.
y= minimum of average distance of that point to other points of other clusters.
Then the Silhouette width, Sil, for a particular point can be defined as,

$$Sil = \frac{(x-y)}{max(x,y)} \tag{1}$$

Silhouette index is computed as average of Silhouette width of all data points of a given data set. It can vary from -1 to 1 and a good clustering solution posses higher Silhouette index value.

B. DB index

Davies-Bouldin index or DB index [2] is an internal cluster validity index to measure the goodness of the formed clusters. It is expressed as,
(Summation of within-cluster separation / the between-cluster separation)
Separation within the i^{th} cluster S_i is calculated as $S_i = [\sum_{\overline{x} \in C_i} d_e(\overline{x}, \overline{c}_i)]/n_i$ where n_i denotes the number of points present in cluster C_i, and $d_e(\overline{x}, \overline{c}_i)$ is the Euclidean distance between the point $\overline{x}$ and the center $\overline{c}_i$ of the i^{th} cluster, C_i. The distance between clusters C_i and C_j, denoted by d_{ij}, is defined as $d_{ij} = d_e(\overline{c}_i, \overline{c}_j)$. Then, DB index is defined as,

$$DB = \frac{\sum_{i=1}^{K} R_i}{K} \quad (2)$$

here,

$$R_i = max_{j,j \neq i}(S_i + S_j)/d_{ij}.$$

If the obtained clusters are compact in shape and also well separated from each other then the DB index value will be the lowest. A good clustering solution corresponds to the minimum value of DB-index.

III. Preprocessing of datasets

Before applying our proposed clustering algorithm on miRNA datasets we have done some preprocessing to normalize the data. As the raw microRNA datasets contain numerical attributes of both high and low ranges, in order to remove the bias-ness at first all expression values are log transformed. Then each miRNA is normalized to have mean 0 and variance 1. All of the proposed experiments are performed on the normalized miRNA datasets.

IV. Initializing input parameters for proposed algorithm

The parameter combinations used for *Auto-MODE* are as follows: population size: 100, maximum number of generations: 30, CR = 0.04 and F = 0.8. The parameter combination is derived after conducting a thorough sensitivity study. As the DE algorithm is highly sensitive to it's parameters, therefore a thorough sensitivity study has been performed to determine the appropriate values of different parameters. During sensitivity study, the following ranges are kept for different parameters:

- number of generations = {20, 30, 40, 50}.
- CR is varied in the range of 0.1 to 1.
- F is varied between 0.5 to 1.

Finally the combination of parameter values for which the best results are obtained are reported above. The proposed approach is executed 10 times on each dataset. The average values of the best results obtained are reported in Table 1 of main manuscript.

V. Proposed entire Process

The basic steps of the proposed algorithm, *Auto-MODE*, are furnished below:

1) Create initial population of size N by randomly generating *Weights* and *Cluster Centers* vectors of individual chromosome (as described in Section II.B of main manuscript).
2) Run FCM algorithm five times on each chromosome to identify the initial partitionings.
3) For each vector, two objective functions, XB and I indices are calculated (as described in Section II.D of main manuscript).
4) Perform non-dominated sorting and crowding distance operations [3] on the solutions of current population.
5) Perform mutation operation on *Weights* and *Cluster Centers* components of the individual chromosomes of the population.
6) Perform classical crossover operation on *Weights* and *Cluster Centers* components of the individual chromosomes of the population.
7) The XB and I index based objective measures are calculated for all the chromosomes of the new population of size N.
8) Merge two populations : parent population(N) and offspring population(N) to get a population of size 2N.
9) Perform non-dominated sorting and crowding distance operations to rank merged population.
10) Apply selection operation to select best N number of chromosomes from the merged population of size $2N$.
11) Repeat the steps 5-10 until the maximum number of generations is reached.
12) Measure the values of Silhouette index [1] for all the generated non-dominated solutions on the final population.
13) Select the solution with highest value of Silhouette index [1].

VI. Complexity analysis of proposed *Auto-MODE* algorithm

In this section the time complexity of the proposed *Auto-MODE* clustering algorithm is analyzed. Let us assume that: n: number of points and d: number of features, $MaxLen$: maximum number of clusters, N: total population size, XB and I are the times taken to compute the XB and I index of a single chromosome, respectively, $tolIndex$: total number of objective functions, $maxgen$: number of generations, $chromLength$: length of chromosome which is $(MaxLen \times d) + d$.

Below we have analyzed the complexities of different steps of the proposed algorithm:

1) Initialization of population takes $O(N \times chromLength)$.
2) Membership calculation for one chromosome takes $O(n \times MaxLen)$ time. For the whole population the total time required for this step is $O(N \times n \times MaxLen)$.
3) Calculation of objective functions for all chromosomes of population takes $O(N \times tolIndex \times (XB + I))$.
4) Non-dominated sorting takes $O(tolIndex \times (2N)^2)$.
5) Crowding distance assignment takes $O(tolIndex \times 2N\ log(2N))$.
6) Selection step of *Auto-MODE* needs $O(N \times chromLength)$ time.
7) Mutation and crossover operations require $O(N \times chromLength)$.

Summarizing the above complexities, overall complexity of the *Auto-MODE* algorithm becomes $O(N \times tolIndex \times (XB + I))$. As, the computational time of XB / I index is $O(n^2)$, therefore the time complexity of proposed *Auto-MODE* becomes $O(n^2 N)$. For maximum $maxgen$ number of generations the total time complexity becomes $O(n^2 N \times maxgen)$.

VII. Obtained Pareto optimal fronts

In order to show the set of final trade-off/non-dominated solutions, the final Pareto front obtained by our proposed approach *Auto-MODE* for three datasets are plotted. These

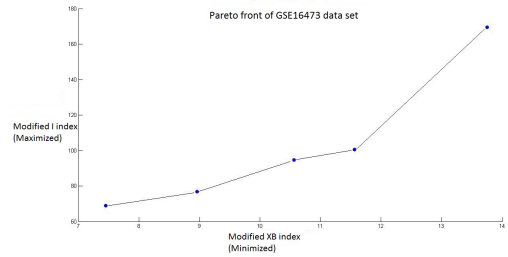

Fig. 2: Pareto front of obtained five solutions for GSE16473 data set

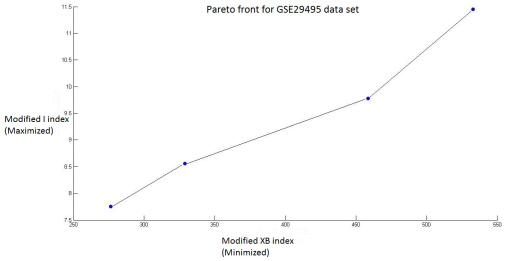

Fig. 3: Pareto front of obtained four solutions for GSE29495 data set

are shown in Figures 2, 3, 4, respectively. Each of the points on the final Pareto front represents one complete clustering solution. Each of these non-dominated solutions corresponds to a complete assignment of all data-points of chosen data set to different clusters. For a particular data set all solutions on Pareto front are non-dominating to each other and are equally good. These figures demonstrate the conflicting nature of the chosen objective functions.

VIII. Biologiccal significance test results for four rest clusters from GSE29495 dataset

The outcomes of biological significance test for four rest obtained clusters for GSE29495 dataset are shown in Table I, II, III, IV.

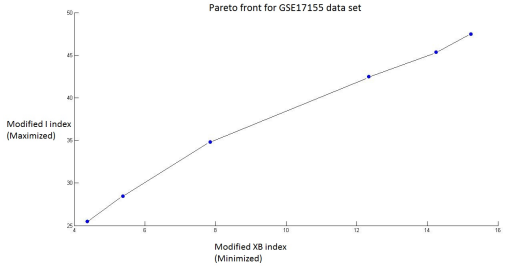

Fig. 4: Pareto front of obtained six solutions for GSE17155 data set

GO term	Module %	Genome %
regulation of cellular process: GO:0050794	77%	46%
macromolecule metabolic process : GO:0043170	69%	27%
regulation of macromolecule biosynthetic process: GO:0010556	57%	16%
single-organism process : GO:0044699	91%	54%
macromolecule metabolic process : GO:0043170	47%	27%
cellular component organization or biogenesis: GO:0071840	54%	21%
response to chemical: GO:0042221	43%	14%

TABLE I: Significant GO terms shared by genes of cluster 2 for GSE29495 dataset

GO term	Module %	Genome %
regulation of primary metabolic process: GO:0080090	43%	23%
single-organism process: GO:0044699	54%	20%
cellular response to stimulus: GO:0051716	65%	27%
nitrogen compound metabolic process: GO:0006807	52%	30%
developmental process: GO:0032502	47%	23%
single-organism developmental process: GO:0044767	47%	23%
single-organism cellular process: GO:0044763	68%	38%

TABLE II: Significant GO terms shared by genes of cluster 3 for GSE29495 dataset

GO term	Module %	Genome %
regulation of biological quality: GO:0065008	48%	15%
regulation of cell communication: GO:0010646	54%	12%
positive regulation of cellular process: GO:0048522	52%	21%
localization: GO:0051179	44%	20%
establishment of localization: GO:0051234	45%	16%
cellular macromolecule metabolic process: GO:0044260	53%	24%
single organism signaling: GO:0044700	53%	22%
multicellular organismal process: GO:0032501	59%	31%

TABLE III: Significant GO terms shared by genes of cluster 4 for GSE29495 dataset

GO term	Module %	Genome %
anatomical structure development: GO:0048856	54%	21%
regulation of primary metabolic process: GO:0080090	64%	23%
anatomical structure morphogenesis: GO:0009653	38%	9%
animal organ development: GO:0048513	45%	12%
response to chemical: GO:0042221	43%	14%
positive regulation of metabolic process: GO:0009893	38%	14%
regulation of macromolecule metabolic process: GO:0060255	44%	23%
regulation of nucleobase-containing compound metabolic process: GO:0019219	33%	16%
regulation of nucleic acid-templated transcription: GO:1903506	31%	13%

TABLE IV: Significant GO terms shared by genes of cluster 5 for GSE29495 dataset

References

[1] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied*

Mathematics, vol. 20, pp. 53–65, 1987.
[2] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, pp. 224–227, 1979.
[3] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182–197, 2002.